IJCSIS Vol. 16 No. 9, September 2018
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

Indexing Service

IJCSIS has been indexed by several world class databases, for more information, please access the following links:

Global Impact Factor
http://globalimpactfactor.com/
Google Scholar
http://scholar.google.com/
CrossRef
http://www.crossref.org/
Microsoft Academic Search
http://academic.research.microsoft.com/
IndexCopernicus
http://journals.indexcopernicus.com/
IET Inspec
http://www.theiet.org/resources/inspec/
EBSCO
http://www.ebscohost.com/
JournalSeek
http://journalseek.net
Ulrich
http://ulrichsweb.serialssolutions.com/
WordCat
http://www.worldcat.org
Academic Journals Database
http://www.journaldatabase.org/
Stanford University Libraries
http://searchworks.stanford.edu/
Harvard Library
http://discovery.lib.harvard.edu/?itemid=|library/m/aleph|012618581
UniSA Library
http://www.library.unisa.edu.au/
ProQuest
http://www.proquest.co.uk

Zeitschriftendatenbank (ZDB)
http://dispatch.opac.d-nb.de/

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2018 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies, IoT
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

*It is our great pleasure to present the **September 2018 issue** (Volume **16** Number **9**) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and digital technologies. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 11800 times and this journal is experiencing steady and healthy growth. Google statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is already indexed in some major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, LinkedIn, Academia.edu and EBSCO among others.*

A reputed & professional journal has a dedicated editorial team of editors and reviewers. On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers & editors for their outstanding efforts to meticulously review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing or reading papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status, making sure we deliver high-quality content to our readers in a timely fashion.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." We would like to thank you, the authors and readers, the content providers and consumers, who have made this journal the best possible.

*For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 16, No. 9, September 2018 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	**Dr Riktesh Srivastava [Profile]** Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	**Dr. Jianguo Ding [Profile]** Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran	**Dr. Naseer Alquraishi [Profile]** University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	**Dr. Kai Cong [Profile]** Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	**Dr. Omar A. Alzubi [Profile]** Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	**Dr. Jorge A. Ruiz-Vanoye [Profile]** Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	**Prof. Ning Xu,** Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	**Dr . Bilal Alatas [Profile]** Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	**Dr. Ioannis V. Koskosas,** University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	**Dr Venu Kuthadi [Profile]** University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	**Dr. Zhihan lv [Profile]** Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	**Prof. Ghulam Qasim [Profile]** University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	**Prof. Dr. Maqbool Uddin Shaikh [Profile]** Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	**Dr. Musa Peker [Profile]** Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	**Dr. Wencan Luo [Profile]** University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	**Dr. Ijaz Ali Shoukat [Profile]** King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	**Dr. Yilun Shang [Profile]** Tongji University, Shanghai, China
Dr. Sedat Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

7. PaperID 31081813: A Novel Modernization Approach for Migration to Native Cloud Application: A Case Study (pp. 49-57)

Kh. Sabiri, F. Benabbou, M.A. Hanine, A. Khammal & Kh. Akodadi

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

8. PaperID 31081815: A Novel Item Recommender for Mobile Plans (pp. 58-62)

Neetu Singh, dept. of Computer Science, Mody University of Science and Technology, Sikar, India
V.K. Jain, School of Engineering and Technology, Mody University of Science and Technology, Sikar, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

9. PaperID 31081818: Mathematical Knowledge Interpretation from Heterogeneous Data Sources– Industry Perspective (pp. 63-67)

Savitha C. (1), Suresh Babu GOLLA (2),
(1) Suresh Babu Golla, Savitha Chinnappareddy, Emerging Technologies and Solutions, Exilant a QuEST Global Company, Bangalore, India
(2) Emerging Technologies and Solutions, EXILANT Technologies, a QuEST Global Company, Bangalore, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

10. PaperID 31081820: The Modeling and Simulation of Wireless Campus Network (pp. 68-78)

Oyenike Mary Olanrewaju, Computer Science and Information Technology Department, Federal University Dutsinma, Katsina State, Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

11. PaperID 31081822: Relaxed Context Search over Multiple Structured and Semi-structured Institutional Data (pp. 79-89)

Ahmed Elsayed, Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt
Ahmed Sharaf Eldin, Prof. and Dean, Faculty of Information Technology and Computer Science, Sinai University; Faculty of Computers and Information, Helwan University, Cairo, Egypt
Doaa S. Elzanfaly, Informatics and Computer Science, British University in Egypt; Faculty of Computers and Information, Helwan University, Cairo, Egypt
Sherif Kholeif, Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

12. PaperID 31081828: Stock Market Data Analysis and Future Stock Prediction using Neural Network (pp. 90-96)

Tapashi Gosswami, Department of Computer Science and Engineering, Comilla University, Comilla, Bangladesh
Sanjit Kumar Saha, Assistant Professor, Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh
Mahmudul Hasan, Assistant Professor, Department of Computer Science and Engineering, Comilla University, Comilla, Bangladesh

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

13. PaperID 31081840: Iris Segmentation by Using Circular Distribution of Angles in Smartphone Environment (pp. 97-111)

Rana Jassim Mohammed, Taha Mohammad Al Zaidy, Naji Mutar Sahib
Department of Computer Sciences, College of Science, University of Diyala, Diyala, Iraq

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

14. PaperID 31081841: Feature-Relationship Models: A Paradigm for Cross-hierarchy Business Constraints in SPL (pp. 112-124)

Amougou Ngoumou, Department of Computer Science, College of Technology, University of Douala, PO 8698 Douala, Cameroon
Marcel Fouda Ndjodo, Department of Computer Science, Higher Teacher Training College, University of Yaounde I, PO 47 Yaounde, Cameroon

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

15. PaperID 31081843: A Clustering Based Approach to End Price Prediction in Online Auctions (pp. 125-141)

Dr. Preetinder Kaur, Course Convenor – ICT, Western Sydney University International College, Sydney, 100 George Street, Parramatta NSW 2150
Dr. Madhu Goyal, Faculty of Engineering and Information Technology, University of Technology, Sydney, 15 Broadway, Ultimo NSW 2007

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

16. PaperID 31081844: Thinging Ethics for Software Engineers (pp. 142-152)

Sabah S. Al-Fedaghi, Computer Engineering Department, Kuwait University, City, Kuwait

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

17. PaperID 31081831: Critical Analysis of High Performance Computing (HPC) (pp. 153-163)

Misbah Nazir, Dileep Kumar, Liaquat Ali Thebo, Syed Naveed Ahmed Jaffari,
Computer System Engineering Department, Mehran University of Engineering & Technology, Sindh, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

New approach for generating frequent item sets without using Minimum Support Threshold

Sachin Sharma *(Research Scholar)*[1], Dr Shaveta Bhatia *(Supervisor)*[1]

[1]*Faculty of Computer Applications,*
Manav Rachna International Institute of Research and Studies,
Faridabad
sachin.fca@mriu.edu.in, shaveta.fca@mriu.edu.in

***Abstract*- Association Mining is preeminent widespread model in Data Mining. In various association algorithms like Apriori, IP-Apriori and Frequent Pattern-Growth, the concept of minimum threshold is used. Major problem in existing algorithm is that customer must define the minimum support threshold. Supposing any customer desires to put on any association algorithm with billions of communication or transactions, and customer does not have the essential information about the transactions, and for that reason, a customer won't be able to decide correct support value. Here, a new system is proposed in which the existing algorithm would stipulate the minimum support value in a completely programmed way. In the proposed system, some mathematical methods have been used for calculating the minimum support threshold value automatically.**

Keywords: Frequent, Threshold, Support, Apriori, FP-Growth

I. INTRODUCTION

Data mining has been energetic in terms of determining patterns from huge set of databank. It has been devised as an intermediary phase of information finding of facts. Data mining can be demarcated as an interdisciplinary process for mapping data sets and to the procedure of visualizing it. It is a method of changing raw data into its logical arrangement. Data mining targets at analysing data into set of data sets; characterised as clusters; or as a set of rare dependencies. The set of dependencies that experience between any groups of data is termed as a procedure of association.

Association has been an important challenge for examining dependency of one data item on the other, which is normally related by means of support and confidence. This study has become prominent in many disciplines like market basket analysis, fraud detection etc. It allows customers to study data from several dissimilar angles, classify it, and encapsulate the associations recognised. Theoretically, Data Mining is a method of generating associations with tons of grounds in big databanks.

An Association rule is an inference of form R1→R2; R1 and R2 are sets of items and R1 n R2 = ⌀.The initial method suggested in association rule is Apriori and lots of methods were imitative from it. All of these methods find interesting rules grounded on the minimum thresholds. Minimum Support Threshold help in sinking the search space and avoid the method from running out of memory. Unfortunately, lesser support, bigger amount of rules obtained that are tough to examine. However, bigger support produces effective but uninteresting rules.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Thus, Association Mining technique generates the rules founded on the accurateness in describing Minimum Support Threshold.

To overcome the above constraint, it is proposed to consider the option of emerging an original structure that works without having to determine a support threshold. The projected model in detail is described in segment 3 (Proposed Methodology) after a discussion of earlier work is accessible in segment 2 (Related work). Experimentations based on an operation of the structure and a discussion of the results is open in segment 4(Results). Lastly, conclusion is ready in segment 5.

II. RELATED WORK

Association mining via Apriori, IP-Apriori and FP-Growth algorithm provides noble results but when the size of database rises, the outcome drops whole database is scanned again and again. This method follows a breadth-first search technique to obtain huge itemsets. The concern with these algorithms is that it cannot be functional to extract big databases. Another method is frequent growth algorithm which follows divide-and-conquer technique. This method uses hierarchy of many patterns and calculates the frequent itemsets.

Yan Hu et al [1] anticipated an enhanced method to extract frequent itemsets. During the mining process, the author at times, requisite to deal with big records of candidate itemsets. Though, meanwhile frequent itemsets are ascendant bunged, this is adequate to learn only all frequent itemsets. The arrangement aids to decrease the number of checking and protect the exploration point in time.

Li Juan et al [2], planned a new approach that can translate a database into a hierarchical form behind data pre-processing, and then perform the association mining of the hierarchical form. This method has extra reliability than FP growth method, and preserves whole data for mining; it will not wipe out the lengthy configuration of any transaction, and considerably shrink the non-interesting information.

Khairuzzaman, T. [3] presented the vibrant graphical or hierarchical rearrangement perception. He suggested CP-tree which attains a frequency-descending arrangement with a particular pass and significantly decreases the execution time to generate frequent itemsets. In their process, the authors have settled that CP-tree attains a significant presentation in terms of complete runtime. The simple maintenance and continuous access to complete database is interactive and incremental.

Yaang Q. et al used enhanced Apriori algorithm to discover the correlation rules which provides the significant information for the curriculum. This algorithm wishes to examine the database only one time when generating candidate itemset, it calculates the support count of other candidate itemsets through stating the count of the parallel set, not scanning the database constantly, which saves the time to a great extent.

Saravanan Suba et al [6] provided the main concepts of Association rule and previous existing methods with their effectiveness and limitations. They concluded that Association rule is an interesting problem but still contains some drawbacks such as it scans the database again and again.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Uno, T. et al [7] proposed various methods for generating itemsets. The methods developed can take out frequent itemsets, closed itemsets, and maximal itemsets. These methods are well-organized in finding interesting knowledge.

Lingaraju P. et al [4] implemented the various methods for generating frequent itemsets and others. The authors built a prototype with GUI to demonstrate the efficiency of data mining algorithms. They tested the model on various datasets and found that this framework is much useful as compared to previous ones and can be used for expert decision making.

Wanjun Yu et al.[5] developed new algorithm which recovers the level of existing algorithm by sinking amount of frequent item set produced in pruning process, by applying transaction tag process.

Sharma and Khurana [12] presented a comparative study of various algorithms AIS, SETM, Apriori, AprioriTID and Apriori Hybrid. The authors concluded that Apriori Hybrid algorithm is superior to Apriori and Apriori TID as it decreases the speed and recovers the accurateness. The limitation of these algorithms is number of scans over the database.

III. PROPOSED METHODOLOGY

This paper presents a new algorithm based on structure to recognise frequent itemset. The new procedure switches random minimum support value defined by user with efficient model grounded on mathematical techniques. In this algorithm, Minimum support value is considered grounded upon average value of support count of all transactions. This methodology make this process more contented for somebody non skilled in data mining.
The detailed steps are as follows:

Algorithm: Proposed **Input:** database of transactions **Output:** Frequent item sets in data base

1: Compute, Minimum Support Threshold= $\frac{\sum((\text{support(i)}}{n}$ where n is number of transactions.

2: For each item, count the occurrence from transactional database and add it to the structure after scanning the data base one time only. Place them into different addresses.

3: To discover all the frequent item sets, scan from the largest item sets in the arrangement. Suppose it be k-item set. If there are more than one k-item sets, we judge the next k- item set until getting all the maximum frequent item sets.

4(a): If the k-item set is frequent, then its subsets are frequent (According to the property of Apriori). Include the Set and all the subsets in the frequent-item set table if not included. Go to step 7.
4(b): If the k-item set is not frequent, then produce its (k-1)-item set subsets.

5: Compute support count for all (k-1)-item set subsets (which are not involved in frequent item set table) from the arrangement and not by scanning the database.

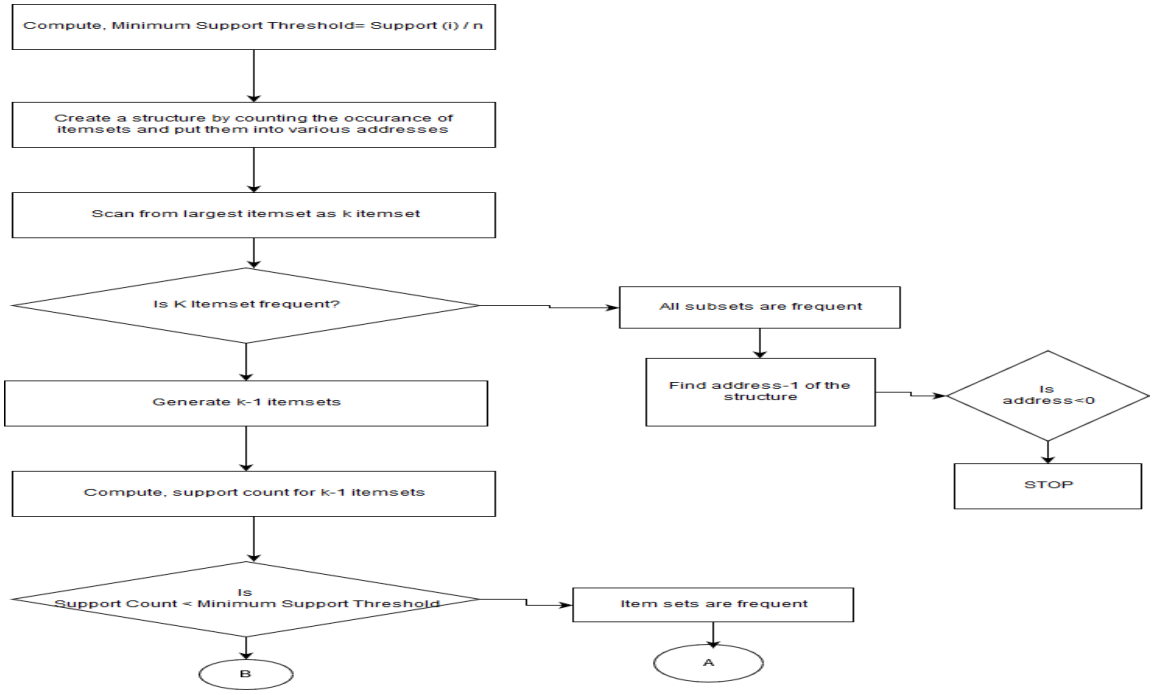
6: Compare each of the support count by min_sup. If the item set is frequent then go to step 4(a). If the item set is not frequent then go to step 4(b).

7: Now, go to address-1 of structure and if address <0 then stop.

8: Scan all the item sets if they are not involved in frequent item set table then estimate the support count for them.

9: Compare each support count with the min_sup. If the item set is frequent then go to step 4(a). If the item set is not frequent then go to step 4(b).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



V. RESULTS

a) Performance Evaluation:

To find the efficacy of projected algorithm, the author has compared the proposed algorithm with existing algorithm respectively. The datasets are executed and tested on machine Intel Core 2, 2.00 Ghz with 64 bit Operating system and are implemented in MATLAB.

The performance of experiment is measured using total execution time and memory usage for generating item set. In this comparison distinct dataset with different threshold support values are considered.

b) Data Set Description:

Various datasets have been used to find the performance of the proposed system. Three datasets Mushroom.txt, Chess.txt and Primary Tumor.txt from UCI Repository library.

TABLE 1
REAL DATASET (CHARACTERISTICS)

Data Set	Instances	Attributes
CHESS	3024	37
MUSHROOM	8124	22
PRIMARY TUMOR	389	1711

These datasets have been attained from UCI (University of California, Irvine) Machine Learning Repository. Each dataset includes diverse instances and attributes. Mushroom billets 22 attributes that is cap shape, surface, colour, class etc. Primary tumor accommodates 17 attributes that is brain, skin, neck, abdominal, liver, age, sex etc. All

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

values of attribute in the database have been arrived as numeric values equivalent to their catalogue in the list of attribute values for that attribute domain. Chess data accommodates 3024 transactions and 37 attributes.

The proposed algorithm is tested on these datasets for various transactions and result obtained are in the form of Minimum Support threshold, number of frequent item sets and time executed.

TABLE 2
RESULTS OBTAINED FOR DATA SET 1(CHESS)

Data Sets	Min Support Threshold	No of Frequent item sets	Execution Time (seconds)
20	4	127	0.1875
200	12	191	5.71
500	23	223	8.01
1000	16	607	117.1
1500	20	671	134.6

TABLE 3
RESULTS OBTAINED FOR DATA SET 2: MUSHROOM

Data Sets	Min Support Threshold	No of Frequent item sets	Execution Time (seconds)
20	1.5	207	0.9
200	6.2	319	6.17
500	12.8	319	6.5
1000	25	319	6.9
1500	15.4	1155	56.1
2000	17.6	647	94
3000	26.3	559	86.9
4000	32.2	479	90.2

TABLE 4
RESULTS OBTAINED FOR DATA SET 3 (PRIMARY TUMOR)

Data Sets	Min Support Threshold	No of Frequent item sets	Execution Time (seconds)
20	1.4	207	1.21
150	5.3	255	1.8
260	9	255	1.82
360	12	255	2.12

-

TABLE 5
COMPARISON BETWEEN PROPOSED AND EXISTING ALGORITHM (CHESS)

	Average No of Frequent item sets	Average Execution Time
Existing Algorithm [Sequeira et.al]	3839	2
Proposed Algorithm	127	0.1875

TABLE 6
COMPARISON BETWEEN PROPOSED AND EXISTING ALGORITHM (MUSHROOM)

	Average No of Frequent item sets	Average Execution Time
Existing Algorithm [Sequeira et.al]	81220	200
Proposed Algorithm	463	14

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 7
COMPARISON BETWEEN PROPOSED AND EXISTING ALGORITHM (PRIMARY TUMOR)

	Average No of Frequent item sets	Average Execution Time
Existing Algorithm [Sinthuaj et. al]	527	93
Proposed Algorithm	243	1.73

Also, the time executed by proposed algorithm is very less as compared to existing algorithms as seen in above mentioned tables. The proposed algorithm scans the database only once which is based on a structure. This structure is constructed by analysing the item sets after scanning the database one time only and put them into different addresses with their support value.

In the proposed algorithm, there is no need to find the support count of various item sets again and again, therefore the number of candidate item sets are also less which shows that memory space is minimised.

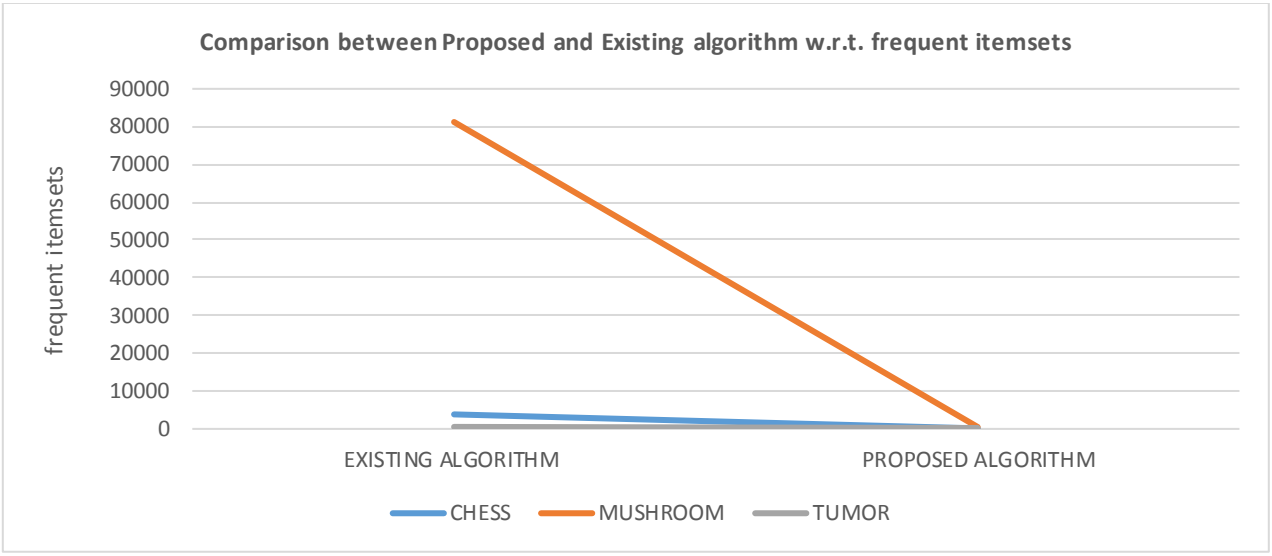


Figure 1: Comparison between proposed & existing w.r.t. frequent itemsets

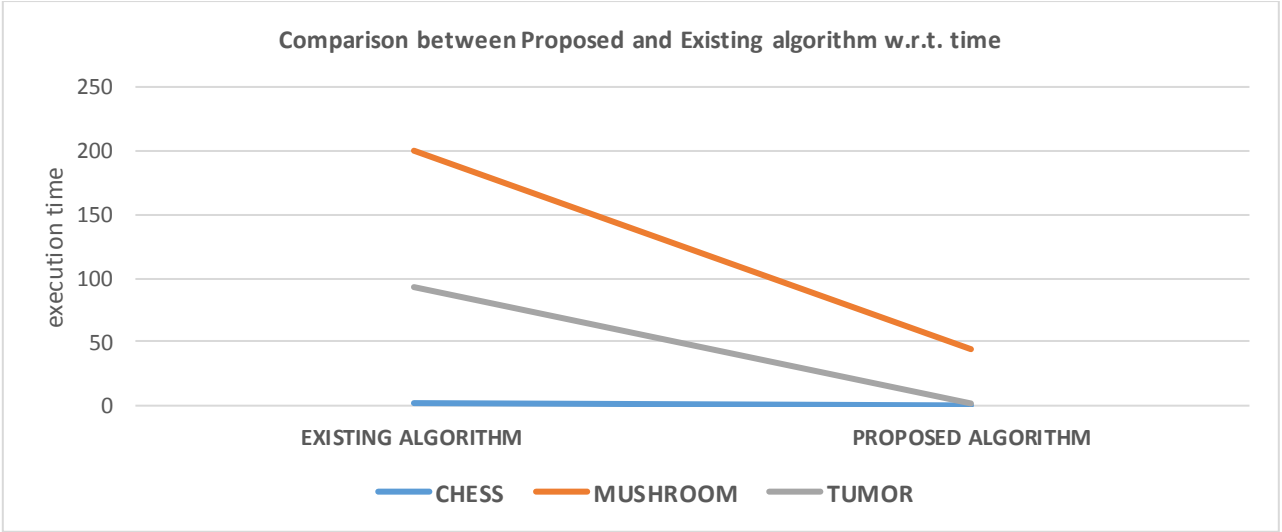


Figure 2: Comparison between proposed & existing w.r.t. time

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The proposed system is executed by calculating the minimum support threshold value automatically whereas the Apriori algorithm uses the random value entered by user.

TABLE 8
RESULTS OF CHESS DATASET

Support	10%	20%	40%	Avg Support
No of frequent itemsets	71807	26495	3967	127
Execution time	156	25	2	0.1875

TABLE 9
RESULTS OF MUSHROOM DATASET

Support	10%	20%	40%	Avg Support
No of frequent itemsets	101989	18799	773	463
Execution time	312	14	0.25	15

Comparison based on frequent item sets generated for support is shown in Fig.3. Data Sets generate high frequent item sets when support is 10%. When support threshold is 20% or 40%, the number of frequent item sets generated are minimised but for average support, frequent item sets generated are very less.

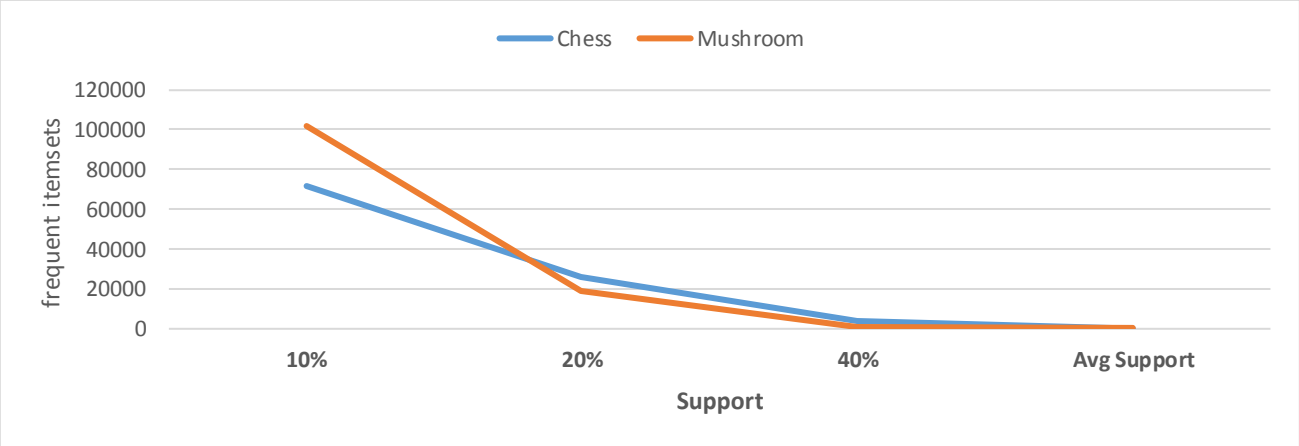


Figure 3: Comparison between proposed & existing w.r.t. frequent itemsets

Comparison based on time is shown in Fig.4. The datasets takes high time when support is 10%. When support threshold is 20% or 40%, time taken is very less but for average support time consumed is neither high nor less.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

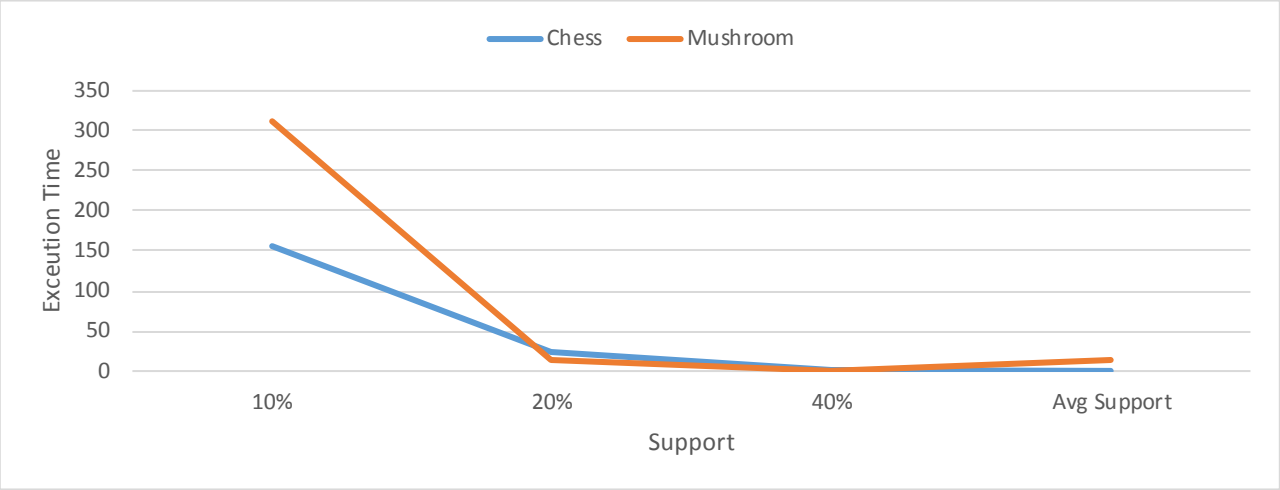


Figure 4: Comparison between proposed & existing w.r.t. time

VI. CONCLUSION

The existing algorithm produces unnecessary candidate item sets and consumes much space. The proposed algorithm takes minimum memory space to produce frequent item sets. There is no need of scanning the database again and again and scans the database once to maintain the structure. Support value is determined using this structure.

In the existing algorithms, a customer may modify support threshold value to search out the enhanced outcome. But the proposed algorithm uses some mathematical functions. Various experiments have been conducted on different datasets; representing Apriori algorithm still leads to a conflict results due to user defined support value. But, after applying mathematical functions during pruning, it produces superior results. Also, the existing algorithm needs lots of time to generate frequent item sets, but analysis on proposed algorithm shows that much time is saved for generating the frequent item sets.

REFERENCES

[1]. Yan Hu, "An Improved Algorithm for Mining Maximal Frequent Patterns", In Proceedings of Artificial Intelligence, IEEE, 2009
[2] Li Juan, De-ting M, "Research of An Association Rule Mining Algorithm Based on FP tree", In Proceedings of Intelligent Computing and Intelligent Systems, IEEE, 2010.
[3] Tanbeer S. Khairuzzaman,"Efficient single-pass frequent pattern mining using a prefix-tree", Elsevier, Information Sciences, 559-583, 2008.
[4] Lingaraju P, "Efficient Data mining algorithms for mining frequent/closed/ maximal item sets", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, No 9, 616-621, 2013[5] W. Yu and X. Wang, "The Research of Improved Apriori Algorithm for Mining Association Rules", In Proceedings of 11th IEEE International Conference on Communication Technology, 513-516, 2008
[6] Saravanan S. "A study on milestones of association rule mining algorithms in large databases", International Journal of Computer Applications, Vol 47, No 3, 12-19, 2012
[7] Uno, T, M. Kiyomi, H. Arimura. "LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Item sets".
[8] Agrawal R and R. Srikant, "Fast algorithms for mining association rules", In proceedings of the 20th VLDB Conference Santiago, Chile, 1994
[9] R. Agrawal, H. Mannila, R. Srikant, H. Toivonenv, A. I. Verkamo, "Fast Discovery of Association Rules.", In Advances in Knowledge Discovery and Data Mining, MIT Press, 1996
[10] J. Han, J. Pei, Y. Yin, "Mining frequent pattern without candidate generation", in Proceeding of ACM SIGMOD International Conference Management of Data, ICMD, 1-12, 2000
[11] Kunkle D, "Mining Frequent Generalized Itemsets and Generalized Association Rules without Redundancy", J. Comput. Sci. & Technology., Vol. 23, No 1, 77-102. 2008
[12] Khurana K and Sharma S, "A Comparative study of association rule mining algorithms", International Journal of Scientific and Research Publications, Volume 3, No 5, 38-45, 2013

AN e-HEALTH PROFILE-BASED ACCESS CONTROL PLATFORM IN A CLOUD COMPUTING ENVIRONMENT.

Idongesit E. Eteng
Department of Computer Science, Faculty of Physical Sciences, University of Calabar
Ekoro, Ekoro Igo
Cross River State College of Education, Akamkpa

ABSTRACT

Background: The lack of a centralized database and proper authentication system for healthcare delivery is a major problem of the paper-based healthcare system in existence across healthcare institutions in Nigeria. The aim of this research is to develop an e-Health cloud application that offers an even distribution of healthcare data among major healthcare stakeholders as well as some level of profile-based authentication in a cloud computing environment. This profile authentication scheme ensures that resources are made available based on the profile of the user.
Methods: The software engineering methodology adopted for this study was the incremental software development model. This allowed the end users experiment with the intermediate version of the system before a final version of the system was developed after their inputs.
Results: The developed system was presented to healthcare personnel, thereafter, a questionnaire testing the usability and acceptability of the system was administered to 75 healthcare personnel. The results of the interview revealed that 86.96% of the doctors, 82.76% of the Nurses, 66.67% of medical laboratory scientists and 63.64% of the medical record staff tested the system gave satisfactory reports on the system performance regarding the usability, functionality , privileges and acceptability of the system. The developed system was recommended for adoption in all healthcare institutions across the state in the first instance, and then the country as a whole.
Keywords: Profile-authentication, e-Health, cloud computing

INTRODUCTION

The term Cloud computing is a word use in describing the provision of computing as a model of services over an internet network. Usually, a customer or an enterprise does not claim ownership of the resources or the mechanism on which the services are domicile, rather they pay premium for the services provided in form of a utility. A simple example is the power transmission grid. Subscribers of electricity do not necessary need to be owners of the power generating plants where the electricity is generated from, nor does the customer owns the plants transmission lines where the power is been transmitted. Customer only subscribes for the quantity of power consumed over a given time frame [1]. A cloud computing framework operates in a similar fashion like a utility firm when applied to an internet environment. Rather than distributing electricity, cloud computing enterprises provide computing resources in the form of data storage and software applications. The service provider company's resources are domiciled on their dedicated servers in a remote and distant area. A customer is at liberty to request and use the computing resources on their individual computer device, these devices can be a desktop computer, a laptop, a tablet or even smart mobile phones.

In a profile-based authentication system, resources and services are available to users based on some predefined conditions referred to as the profile of the user. In this system, resources not prescribed for a particular user are not accessible to such a user. The major advantage of the profile-based access control system is that the entire system is compartmentalized into several integrating and overlapping units which ensures privacy of some confidential data.

As a result of technological developments, conventional healthcare practices have greatly been influenced. As a result, the health sector has migrated from

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

hitherto situations where paper-based medical prescriptions was mostly used to Personal Health Records (PHR), Electronic Medical Records (EMR) and the recently Electronic Health Records (EHR)[2]. The necessity to incorporate patients' medical information from distant and remote area, such as primary healthcare institutions, secondary healthcare institutions, tertiary healthcare institutions, clinical laboratories, health insurance organizations and non-governmental organizations has metamorphosed to the technology referred to e-Health. The World Health Organization (WHO) in its definition deduced that e-Health refers to the transfer and communication of healthcare resources to various healthcare practitioners and clients by using the emerging Information Technology (IT) equipment and electronic commerce procedures [3]

Nonetheless, the interchange and incorporation of electronic medical data domiciled and control by various healthcare practitioners and other integrating enterprises is overblown and tasking to control, which inevitably demands for the deploying of cloud computing services in the healthcare domain [2]. The cloud computing (CC) model has assisted the healthcare domain of the challenges hitherto faced by the traditional healthcare systems, the traditional healthcare system uses paper-based medical records (PMR) in storing medical data. Paper-based medical record may be a time-consuming and tedious work, this is because the paper record can only be available in one place at a time, and for this reason, this documents cannot be accessed between two or more medical specialists from different and diverse places concurrently. Also, in the traditional healthcare systems, referral services are done on written documents from a physician to another healthcare institution through the patient or relatives of the patient. This document may also be misplaced or

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

mutilated which can lead to an ineffective referral system. The cloud computing architecture has freed healthcare organizations of the demanding and tedious responsibilities of infrastructure management such as keeping large chunks of paper-based medical records in large file cabinets which may be affected by natural hazards such as floods and termites or by man-made hazards such as fire and theft, hence the necessity to become familiar to third-party Information Technology service providers to forestall this menace becomes inevitable [5]. Moreover, according to [6], the cloud computing model has displayed enormous potentials in enhancing integration among the various collaborating entities of the healthcare realm and to provide the most expected gains, which include scalability, swiftness, cost effectiveness, and all time accessibility of healthcare related data.

Apart from the problem of integration of healthcare services, security and interoperability are major challenges militating against healthcare givers in the discharge of effective healthcare systems. Presently in Nigeria there is no e-Health cloud platform that synergizes the activities of healthcare delivery at distant and remote area [4].This situation is worrisome because challenges and success at a point cannot be accessed from remote locations.

According to [3], the e-Health Cloud computing is the modern computing paradigm that creates an interactive platform for healthcare services. This platform provides a synergy for various healthcare providers at distant and remote areas to interact and access any breakthroughs in healthcare. It is therefore on this assumption that this research intends to deploy an e-Health system with a profile-based access which provides different levels of privileges and authentication to various stakeholders who are involved in the provision of healthcare services. This platform ensures

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

effective distribution of healthcare data (such as patients' records) and services (such as referral services) across major healthcare institutions. In this work, an e- Health platform will be created. The information in this e-Health platform will also create a repository for doctors and their areas of specialization in a particular healthcare institution, this will enable patients to locate doctors, it will also include a referral system where patients can be transferred from one healthcare institution to another based on available facilities. Major stakeholders such as patients, doctors, and nurses will have a platform for robust interaction.

METHODS

This research follows an incremental software development beginning from specification to validation of the developed system. Incremental development is a paradigm that relies on the concept of developing an early implementation, revealing the developed system to end users' comments and rebuilding it iteratively through several versions till a suitable acceptable system has been developed. One major characteristics of the incremental model is that system specification, system development and system validation process are interwoven rather than separate, with speedy feedback across activities.

Incremental software development was adopted for this research because the developed system requires inputs from the users at each increment (modules). In this software engineering model, each version of the system contains some of the basic functionality that is of utmost need by the user. In general, the initial version of the developed system contains the most crucial required functionality. In this methodology, the user can experiment with the system at an early stage in the system development, this will ascertain if the system requirements are met.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

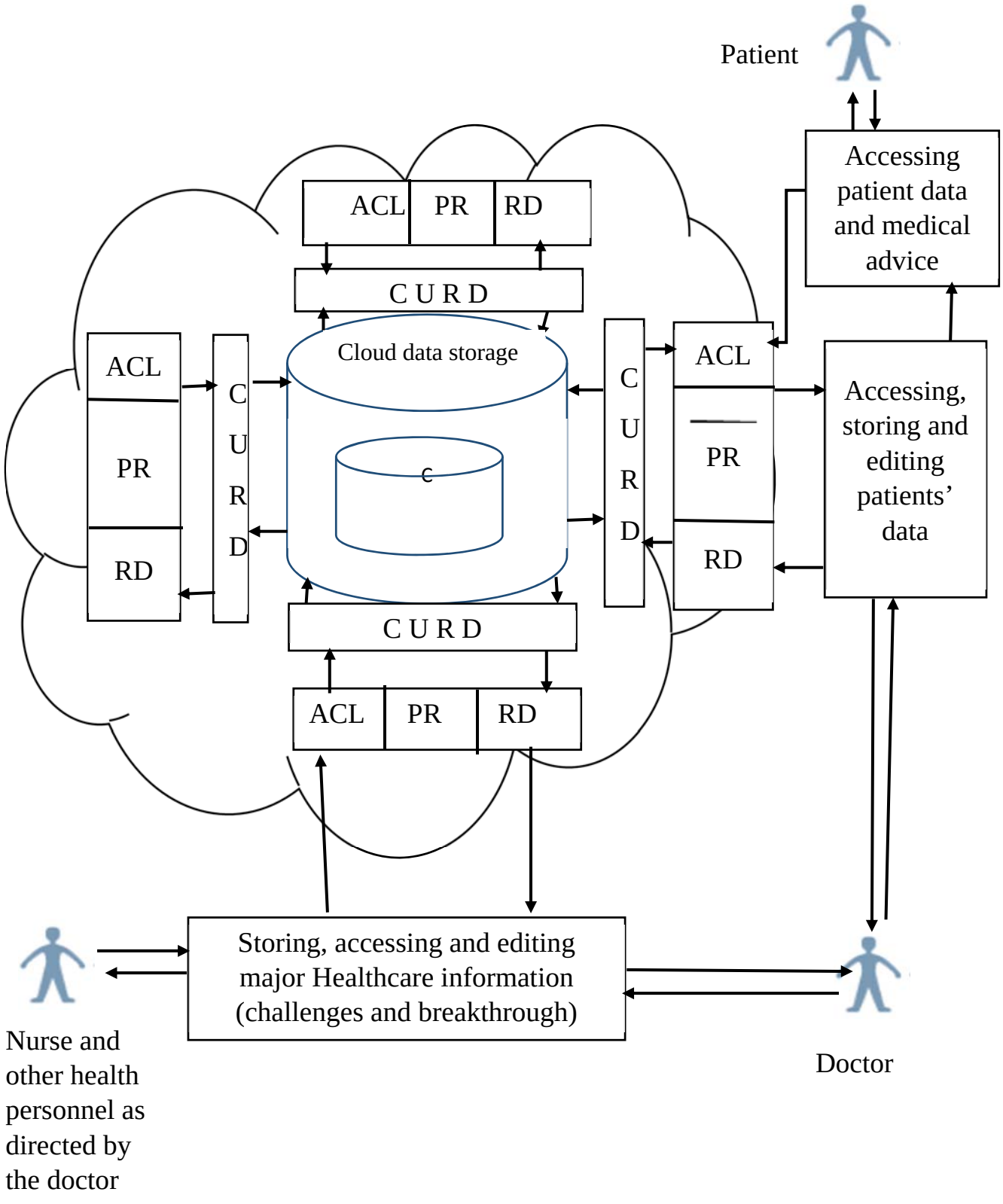


FIGURE 1: System Architecture

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The architecture in figure 1 illustrate the profile authentication of an e-health system in a cloud computing environment. The access control list (ACL), define the resources accessible to
a particular profile. These profiles defines the various health care providers in the platform. The profile gateway (PR) defines access policies and resources in the platform. The resources are mainly the patients' information which are categorized and made available based on the profile of the user. The rule dictionary (RD) specify who perform complete or partial CURD operation. Create, update, retrieved and delete (CURD) operations are privilege operations assigned to users of the platform to restrain some users from performing delicate operations. Some users are privileged to perform complete CURD, while others can only perform partial CURD. The cloud data storage in the platform creates a repository for all data generated in the platform. Another important feature of this architecture is that the doctor is placed at the centre of activity. The activities of the nurses and other healthcare practitioners can easily be monitored by the doctor. Patients can initiate consultations with doctors based on their area of specialties. Doctors in turn can also schedule and reschedule patients' visitations, this invariably reduced the bottleneck associated with the paper-based medical system. This architecture enables patients to access medical practitioners based on their area of specialties. This architecture also creates a conglomerates of medical communities, several healthcare practitioners interacts and cross fertilize ideas that improve the delivery of healthcare services to clients in a more convenient manner than the paper-based healthcare services. With the CURD operations properly spelt out, patients can properly access their medical history without any alterations.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

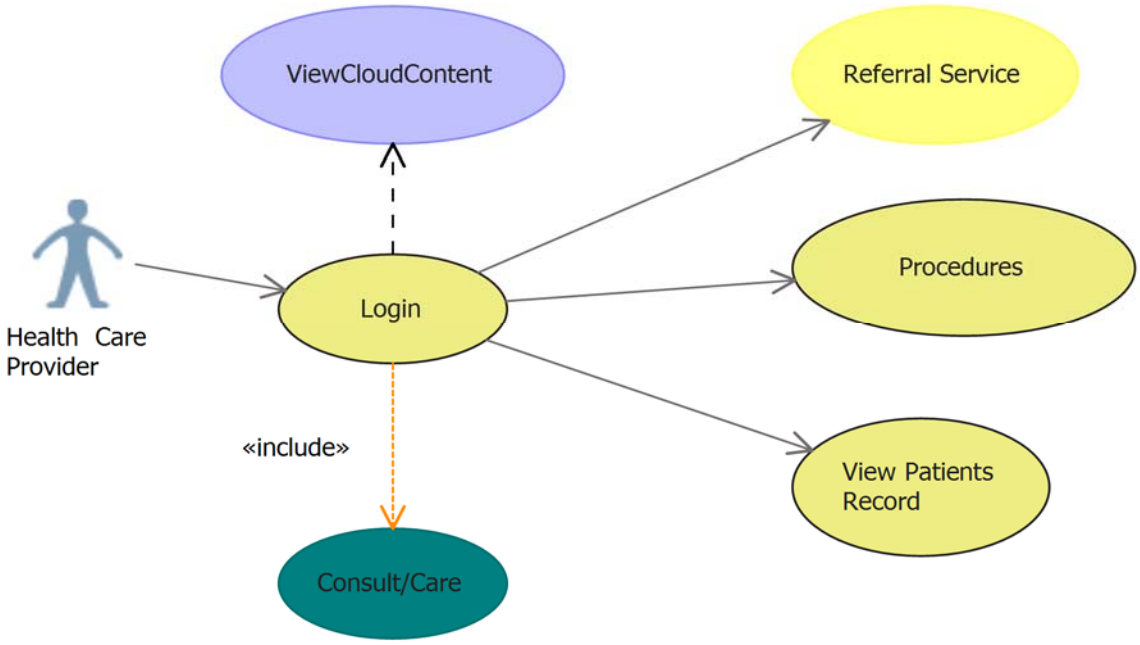


FIGURE 2: Use case diagram for login process of Healthcare providers.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

As illustrated in figure2, the use-case diagram of the healthcare provider demonstrates the activities that exists in linking patients' record among several healthcare institutions which eventually facilitates referral services. This interaction reduced the bottleneck encountered by healthcare providers in getting patients' medical history.

In the use case diagram, health providers are authenticated into the platform through a login screen which assign resources to the user based on the profile of the users, patients' medical records can be accessed.

SOFTWARE REQUIREMENTS

The system was developed using several development tools, these include:

i. Operating system: window 7 was the operating system used in the development of the platform. This was because its powerful performance and friendly user interface.

ii. HTML5: this is the latest version of Hypertext Markup Language.

iii. Cascading style sheet (CSS3): this enhances the presentation of website it is noted for its aesthetic value in web development. Combining HTML5 and CSS3 improves the measure of for a responsive web design technique.

iv. Firefox web browser

v. Tomcat server version 7.0.50

vi. MySQL Database server Version 5.0

vii. Java programming Language

Hardware requirement.

The platform was developed in an HP system with the following configurations

i. Hard disk capacity: 500gigabyte

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ii. Processor: Intel Xeon Processor

iii. Ram size: 4GB

Networking requirement.

a) CAT5 RJ45 cable

b) RJ45 connector

c) Crimper

d) 8-port D-link switch

System deployment

The developed system was tested by deploying the platform to a window server 2012 operating system, this was the server application that acted as the private cloud server. Other clients were linked up with the server for sharing of system resources.

System implementation and testing

The system was implemented and tested on four personal computers and one server meeting the minimum hardware and software requirements with Internet facilities. When the application was launched, the login interface pops up. The clinicians and patients seeking authorization input their username and password in order to gain access and access resources in the system. When an authorized patient's and doctor's correct login details are entered in the username and password columns, the system compares the details with those registered for the patients and clinician in the database and the system grants access. A few test were used in validating the system. The tests are briefly explain below.

Unit Test

Several unit tests were conducted on the different modules of the system. These unit tests includes:

- **Database unit tests:** this test were carried out to ensure that the program was able to connect properly with the MySQL database server,
- **User interface unit test:** these tests were carried out on all the user interfaces in order to ensure that all elements of the user interface were

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

properly displayed at their expected locations.

- **Profile unit tests:** this test was carried out to ensure that users only access resources appropriated to their profiles.
- **Data input unit test:** this test was carried out to confirm that the several data entry modules were able to accept data in the expected and correct formats.

Integration test

The e-Health platform was tested incrementally to ensure that all aggregating units and modules interfaced appropriately. All modules from the user login module to the registration and report screen modules were tested to confirm that they all work together as anticipated.

System validation

The developed system was presented to 75 healthcare personnel consisting of 23 medical doctors, 29 nurses, 12 medical laboratory scientists and 11 medical record staff. In the course of using the system, some medical doctors faulted the initial design of the system which made doctors the desk officers in registering patients. This then prompted the redesign and redevelopment of the system to create room for the front desk officer, thereafter the system was then represented. The response obtained from the healthcare personnel after redesigning the system to accommodate input was a near total approval of the system for our healthcare system in the modern era.

RESULTS

The results of this research shows that the developed system has been tested and approved by healthcare personnel. The result in table 8 shows that 86.96% of the doctors interviewed recommended the software for medical transactions. The introduction of profiling into healthcare database was equally given approval as 78.26% of the doctors accepted that the introduction of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

profiling was a welcome development. Also, 73.91%of the doctors interviewed were impressed by the service provided by the platform. However, only 60.87% of the doctors interviewed agreed that allowing patients access to their medical record was a welcome development.

The results in the responses from nurses that were interviewed shows that 82.76% of the nurses agreed that the software reduces the burden on medical personnel in the management of healthcare information. The confidentiality of medical records introduced by the software was given an enormous approval, this is evidence from the fact that 93.10% of the nurses agreed that the platform improves the confidentiality and safety of healthcare data. The direct interaction of patients with doctors as introduced in the platform was received 82.76% which implies that the platform will reduce the difficulties encountered by patients in the course of seeking medical advice. The result shows that 66.67% of medical laboratory scientists interviewed agreed that the software was interactive and user friendly enough for their usage. The introduction of confidentiality to medical data in the software was given a massive approval as 91.67% of the medical laboratory scientists interviewed agreed that the platform improves the confidentiality and safety of healthcare data. Also, the result obtained from the interview of medical record staff shows a clear acceptance of the developed system. From the result, 100% of medical record staff were impressed with the services provided by the software.

A graphical representation of the analysis is shown in chart 1 and chart 2. Based on the responses obtained and the results shown in the tables and charts from the four categories of responses of respondents, it is deduced that the users are satisfied with the system.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

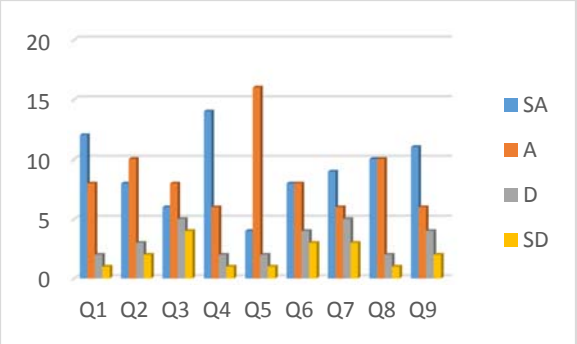


Chart 1: Summary of Doctors evaluation result.

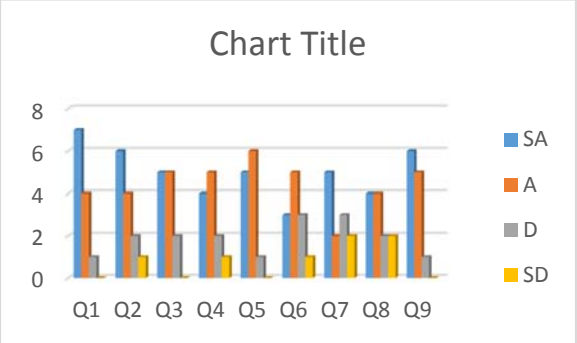


Chart 3: Summary of medical laboratory scientist evaluation result

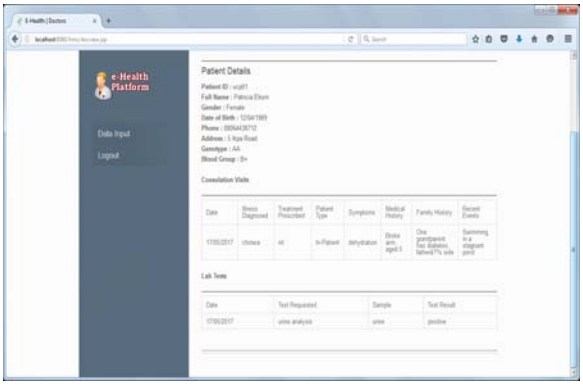


FIG 2: Patient Report Screen (Doctors)

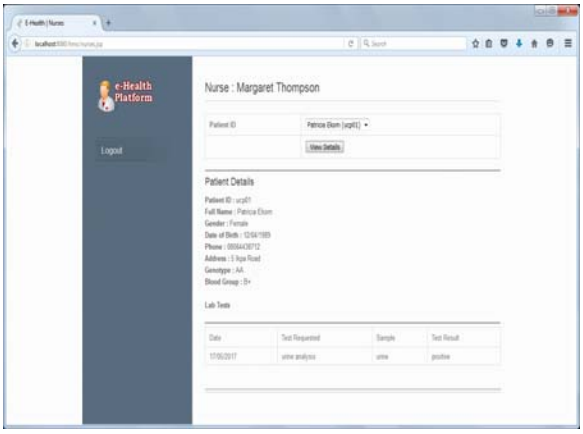


FIG. 3: Patient Report Screen (Nurses)

SUGGESTION FOR FUTURE STUDY

Based on the area of coverage of this work, the following are suggested for future study.

- The SSL/TLS technology be incorporated into the security mechanism of cloud e-Health. This will improve the security architecture of the e-Health platform.
- Biometric authentication scheme in e-health is also another authentication scheme that needs attention in cloud e-Health technology.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

1. Stefan F., Claudia L., & Christoph R. Key Performance Indicators for Cloud Computing SLAs. The Fifth International Conference on Emerging Network Intelligence, 2015. 2(3) 60-64.
2. Zhang, R., Liu, L. Security models and requirements for healthcare application clouds.*3rd IEEE International Conference on Cloud Computing*, Miami, FL, USA, July 2010(268–275).
3. Assad A. & Samee U. K. E-Health cloud: Privacy concerns and mitigation. *IEEE* Journal *of Biomedical and Health Informatics*, 2015. 18(2), 419-429.
4. Edjie E. & Ekabua O. Funding E-Health in Nigeria by NGOS/Multinational Organization: Overview and Perspectives. 2015. International Journal of Computer Applications (0975 – 8887) Volume 111 – No 11, February 2015
5. Abbas, A. K., Bilal, Zhang, L., & Khan, S. U. A cloud based health insurance plan recommendation system: A user centered approach, *Future Generation Computer Systems*, 2015. 44(99-109)
6. Ahuja, P., Sindhu, M. & Jesus, Z. A Survey of the State of Cloud Computing in Healthcare, Network and Communication Technologies, 2012. 1(2): 12-19.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Modified Edge Detection Algorithm Using Thresholds

Jaskaran Singh[*,1], Bhavneet Kaur[2]
[*,1] Department of Computer Science and Engineering, Guru Teg Bahadur Institute of Technology, New Delhi, India.
[2] University Institute of Computing, Chandigarh University, Mohali, Punjab, India.

Abstract- **The edge detection is recognized as a primary technique of computer vision to acquire the actual contour of the object. Various approaches were adopted by researchers to analyse the image edges. Respective method varies on the basis of their parameters and data sets. There are numerous techniques for edge detection such as a first order derivation method, second order derivation method for edge detection, Marr-Hildreth method, Sobel edge detection method, Canny edge detection and many more. Few techniques resulted in some shortcomings such as broken edges and thick boundaries. In this study, an effective approach is proposed for edge detection from the images, where the major focus is on edge steadiness. Both qualitative and quantitative analysis is carried out in comparison to existing techniques. From the simulative outcomes better quality of edges are observed in respect to edge continuity.**

I. INTRODUCTION

Image processing is a field that witnessed the drastic evolution of digital computers. It is a technology which consists of multiple operations like extraction of features, recognition of patterns, segmentation etc. [1]. It is basically a discontinuity in the values of gray level. It has been observed that the majority of information about an image is presented on the object`s contour. Foremost benefit to identify the edges is in the reduction of data amount present in the image which diminish the storage space along with preservation of structural properties of the image. Thus, it is helpful in the extraction of valuable features for pattern recognition [2]. Various categories of edge detection techniques are deeply conversed below:

A. Sobel Edge Detection

Sobel edge detection is a gradient based technique, introduced by Sobel in 1970. Highest gradient points are used as edge points. Basically, it computes the estimated gradient magnitudes on each point of the inputted grayscale image, with an aim to avoid calculated gradient [3][4]. G_a and G_b are implemented using convolution masks:

-1	-2	-1
0	0	0
+1	+2	+1

G_a

- 1	0	-1
- 2	0	+2
- 1	0	+1

G_b

Sobel Edge Detection

// Gradient of function f (a, b)

$$\nabla f = [G_a, G_b]$$

// Magnitude of vector

$$\nabla f = mag(\nabla f)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$mag(\nabla f) = [G_a^2+G_b^2]^{0.5}$$

// Partial derivatives having constant $\phi = 2$

$$G_a = (X_2 + \phi X_3 + X_4) - (X_0 + \phi X_1 + X_6)$$

$$G_b = (X_0 + \phi X_1 + X_2) - (X_6 + \phi X_5 + X_4)$$

// Gradient vector points in the direction of maximum rate change occurred at (a, b). Thus, maximum changed rate angle is:

$$G(a, b) = \arctan(G_b/G_a)$$

B. Prewitt Edge Detection

Prewitt edge detector conveys the indistinguishable constraints similar to Sobel, proposed by Prewitt in 1970. Majorly, it was estimated for eight directions and correspondingly all the eight convolution masks were computed, from these results one with the largest module was selected. Unlike, Sobel edge operator, Prewitt operator avoids emphasis over nearby centre masks pixels. It was observed that the Prewitt edge detector was simpler to implement but produced slightly simpler results in comparison to Sobel edge detector [5][6].

-1	-1	-1
0	0	0
+1	+1	+1

G_a

- 1	0	1
- 1	0	1
- 1	0	1

G_b

C. Canny Edge Detection

The Canny edge detector is a gradient based edge detector, introduced by John Canny in 1983. It identifies the edges from noisy image without disturbing the features of edges [7]. It provides with the best possible results in all possible scenarios. Because of such reason canny edge detectors has been treated as a standard till date [8]. The steps to perform canny edge detection are deeply discussed.

Canny Edge Detection

1) Smoothing of an image is performed using Gaussian filter and to reduce the noise a particular standard deviation σ is specified.
2) Compute gradient magnitude for partial derivatives using finite-difference approximations.
3) Non-Maxima suppression is applied to gradient magnitude.
4) Smoothing is computed for an image $I[\alpha, \beta]$. Gaussian Smoothing Filter is presented as $G_s[\alpha, \beta, \delta]$, where δ depicts the Gaussian spread and manage the smoothness degree.

$$S_m[\alpha, \beta] = G_s[\alpha, \beta, \delta] * I[\alpha, \beta]$$

5) $S_m[\alpha, \beta]$ is smoothed array used to compute α *and* β partial derivatives $A[\alpha, \beta]$ *and* $B[\alpha, \beta]$ as:

$$A[\alpha, \beta] \approx (S_m[\alpha, \beta+1] - S_m[\alpha, \beta] + S_m[\alpha+1, \beta+1] - S_m[\alpha+1, \beta])/2$$

$$B[\alpha, \beta] \approx (S_m[\alpha, \beta] - S_m[\alpha+1, \beta] + S_m[\alpha, \beta+1] - S_m[\alpha+1, \beta+1])/2$$

By averaging finite differences, the computation of α and β partial derivatives are made

6) Gradient Magnitude and orientation are computed as:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$M_g [\alpha, \beta] = \sqrt{A[\alpha,\beta]^2 + B[\alpha,\beta]^2}$

φ [α, β] =arctan (B [α, β], A [α, β]) // arctan (x, y) function used to generate an angle.

Note: Canny Edge Detector estimates the operators that optimizes the outcomes of signal-to-noise ratio and localization.

D. Phase Congmono Detection

Phase congruency is an illumination and feature based edge detector. It detects features from all phase angles [9]. The respective function computes the phase congruency via. monogenic filters [10]. It has average speed and reduced memory requirements as compared to the other phase functions of congruency [11].

II. Prior Work

Evolution in the state-of-the-art begins from 1970s and continued till date. Plentiful algorithms were proposed under the headline. Numerous authors contributed their work over edge detection. In this paper, major contribution in the state-of-the-art from 2010-till date are discussed for better concept clearance. In the year 2010, Wenshuo Gao [12] projected a new edge detection method which used soft-threshold wavelet technique for noise removal from the image and later applied Sobel edge detection operator for desired outcomes. Improved outcomes were observed from multiple experiments on noisy images in comparison with traditional methods.

By the year 2011, Caixia Deng [13] developed a fusion edge detection method which was made with amalgamations of the Prewitt operator, improved Sobel operator, canny algorithm and wavelet transform. The experimental results depict the better noise dealing ability and edge continuity in resultant images. Santa Gupta [3] in the year 2012, presented a fusion edge detection method which were made with combinations of the Prewitt operator, improved Sobel operator, canny algorithm and wavelet transform. The experimental depicts more accurate and effective results. Macro Cal [14] developed a Roberts filter-based edge detection method using CUDA and OpenGL which was executed on the GPU. On the basis of results, high performance and optimized computational time was observed with sequential version of CPU. Later, faster a lines, curves and ellipse detection algorithm were developed by Lianyuan Jiang [15]. Om Prakash [16] proposed an effective modified ABC algorithm for edge detection. The proposed method used pixel position count as a solution. Better entropy values and thick edges were observed from the experimentations. Kavita Sharma [17] proposed an unwanted edge removal method. The color based edge detection was conducted to compute the performance of ACO techniques. B. Gardiner [2] proposed an effective edge computational approach using linear combinations of edge maps at lowest scale. No loss in accuracy and significant edges were identified from the experimentations. Continued progress in the state-of-the-art from the last few years till date are parametric described and presented in Table I.

TABLE I. TABULAR REPRESENTATION OF THE MAJOR CONTRIBUTED PAPERS IN THE FIELD OF EDGE DETECTION.

Author(s)	Algorithm/ Technique	Parameter(s)			Objective(s)	Result(s)	
J. Yu [12]	Improved Sobel Edge Detection Algorithm	Clarity	Threshold	Noise	Noise Removal	Improved Outcomes	NA
Caixia [13]	Improved Edge Detection Algorithm Using Sobel	Clarity	Threshold	Noise	Advantages of Traditional Methods	Accuracy	Effectiveness
S. Gupta [3]	Sobel Edge Detection Algorithm	Noise	Threshold	NA	Work over Noisy Image	Improved Outcomes	NA
Cali M [14]	Performance Analysis of Roberts Edge Detection Using CUDA and OpenGL	Speed	Execution Time	NA	Performance Improvement	Less Execution Time	High Performance
Lianyuan Jiang [15]	Randomize Hough transformation	Speed	Accuracy	NA	Work over Timing and Storage	Better Computation	NA
Om Prakash [16]	An Optimal Edge Detection Using Modified Artificial Bee Colony Algorithm	Redundancy	Threshold	NA	Improved Edge Detection	Better Results	NA
S. Kavita [17]	Ant colony optimization algorithm	Effectiveness	NA	NA	Work over Image`s Color Components	Better Results	NA
B. Gardiner [2]	Multiscale Edge Detection using a Finite Element Framework for Hexagonal Pixel-based Images	Accuracy	Effectiveness	NA	Edge Detection using Linear Combinations	Accuracy	Effectiveness

III. PROPOSED ALGORITHM

Proposed algorithm works on the principle of four basic directions i.e. (north, east, west and south). It uses two threshold intensities (experienced intensities) which, when merged with direction helps in the identification of effective edge points [16].

From the study of various edge detection techniques, some drawbacks have been encountered which are discussed in Table II. To overcome these drawbacks a new optimized edge detector has been proposed in this paper [12] [13] [18].

.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II. DRAWBACKS OF STUDIED EDGE DETECTION TECHNIQUES

Drawbacks	Encountered usually in
Discontinuity in edges	Phase congmono, Prewitt
Thick edges	Phase congmono, Sobel
Detecting useless information.	Canny, Sobel
Processing Speed	Phase congmono
Loss of crucial informative edges	Sobel, Prewitt

A. Flow Chart

Traditional approaches of edge detection majorly use fixed parameters for outcomes, which later doesn't lead to acceptable results. Therefore, to overwhelmed this parametric constraint an algorithm is proposed which later leads to optimized outcomes. The flow chart for the better concept clearance is shown in Fig. 1.

Step 1: Read the image from the database.

Step 2: The respective graythresh is then calculated of the inputted image. The mathematical formula for computing the respective threshold is:

Initialize $\psi = (\sum_{j=0}^{m} \sum_{i=0}^{n} \frac{ara}{m*n})$ (1)

Where ara is an input image array. Thus, the respective equation for evaluating the threshold is presented in Eq. (2)

Threshold value = $\Psi * 0.0039$ (2)

Step 3: On evaluating the threshold the enhancement in the image is perused by adjusting the contrast. The mathematical equation used for it:

If $(\psi > \min (\sum_{j=0}^{m} \sum_{i=0}^{n} ara))$ (3)

$\sum_{j=0}^{m} \sum_{i=0}^{n} ara += \Psi$

Step 4: A later conversion of the image into grayscale is performed if test image is colored using the following equation:

rgb2gray (ara) (4)

Step 5: On the basis of the obtained thresholds effective edge pixels are shortlisted. The mathematical equation to perform the respective activity is as follows:

if (ara [i] [j] > Ψ * ara [i-k] [j-k]), where k= 0,1,-1

$\sum_{j=0}^{m} \sum_{i=0}^{n} out\ [i][j] = 1$ (5)

Where out[i][j] is a resulting array.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

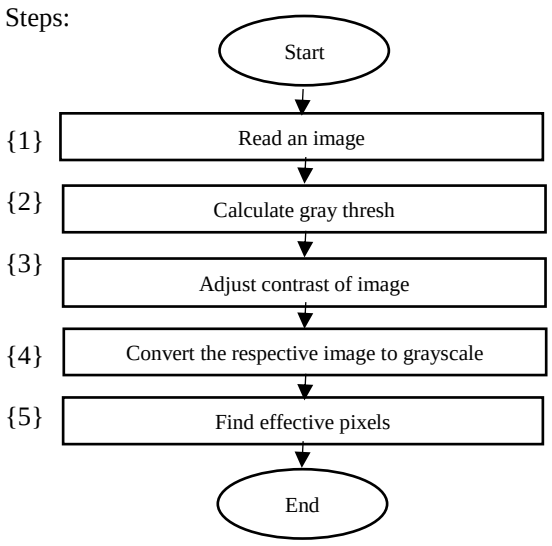


Fig. 1: Flow Chart of the Proposal Algorithm

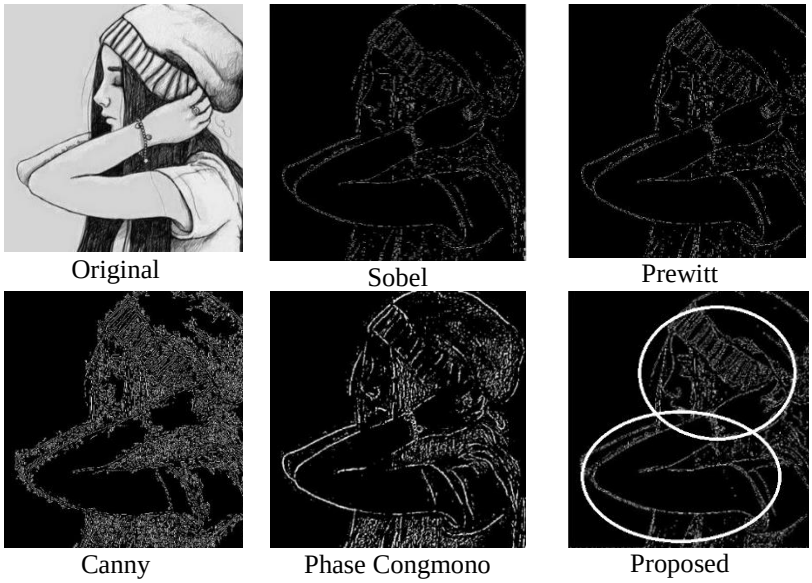


Fig. 2. Comparison on test image 'Girl' using various edge detection techniques.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

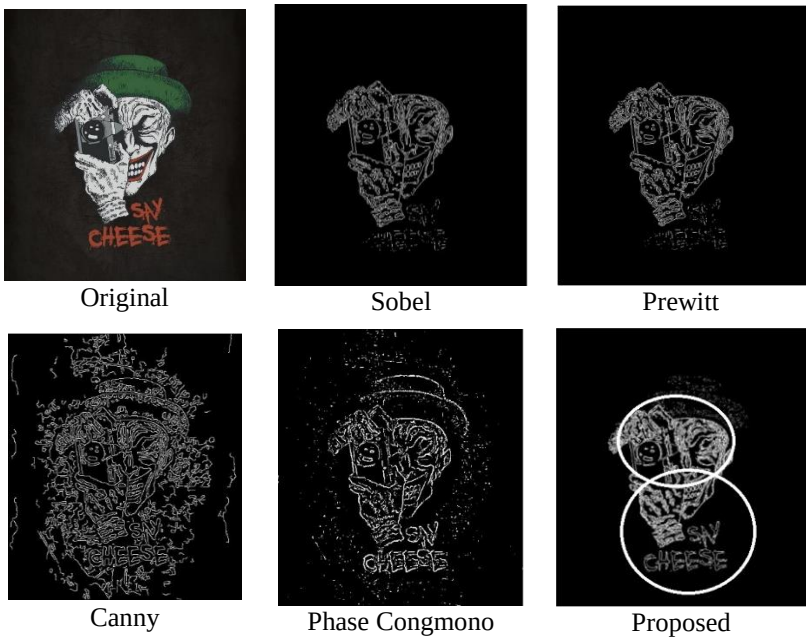


Fig. 3. Comparison on test image 'Clown' using various edge detection techniques.

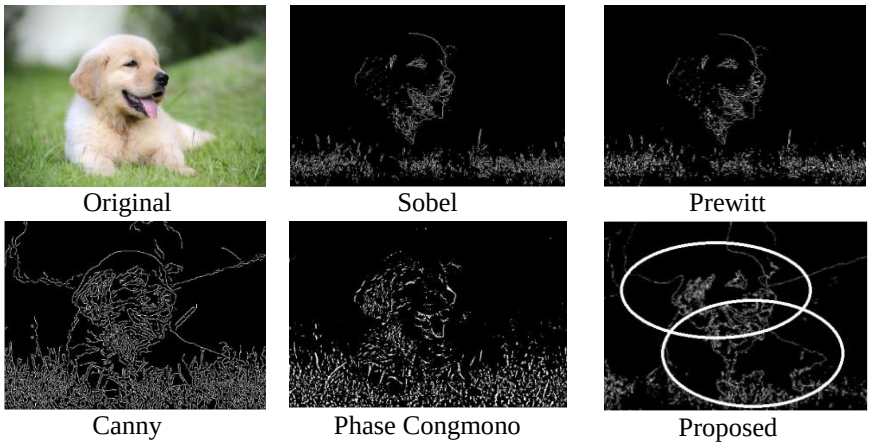


Fig. 4. Comparison on test image 'Animal' using various edge detection techniques.

IV. PERFORMANCE EVALUATION

To test the proposed algorithm, three test images are shortlisted. Girl, Clown and Animal. The results of this study are presented in Fig. 2, 3 and 4. From the experiments, it has been identified that the proposed algorithm results nearby to the most used edge detection technique i.e. Canny. Beside this a comparison with traditional techniques such as Sobel, Prewitt and Canny, along with this attest approach Phase Congmono are made.

It is defined as:

$$H(I) = \text{sum}(c.*\log2(c)) \quad (6)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Here *c* specifies a histogram count and *I* depicts a multidimensional image. Entropy states that too less value of *H* represents the presents of low content and too high value of *H (I)* represents the noise. Thus, for optimal quantitative analysis the values of entropy must be obtained between the upper limit and lower limit. Table III represents the calculated Entropy values of various edge detection techniques over the test images. The graphical analysis of the simulated results in comparison of previously existing methods through entropy is presented in Fig. 5.

V. Results And Discussions

In Table III, it has been identified that Sobel and Prewitt methods always computes a lower amount of entropy count, which illustrates the loss in the crucial informative edges as presented in Fig. 3 and 4. This also leads a lack of continuity in the edges. When there was range of same color shade in an image, it leads to reduction in efficiency of Canny as it detected the meaningless information.

From the Fig. 4, it is identified that the Phase Congmono method detected thicker and discontinues edges to which its entropy value is comparatively high. In contrast, proposed method provided with a continuous edge, least amount of noise and thinner edges as compared to phase congmono in most of the cases; therefore, lied to middle range of entropy values.

TABLE III. ENTROPY VALUE FOR TEST IMAGES OF VARIOUS EDGE DETECTION TECHNIQUES

Images	Sobel	Prewitt	Canny	Phase Congmono	Proposed
Girl	0.1704	0.1699	0.4695	0.3656	0.3475
Clown	0.2072	0.2089	0.457	0.275	0.3066
Animal	0.2061	0.2066	0.4441	0.3328	0.2581

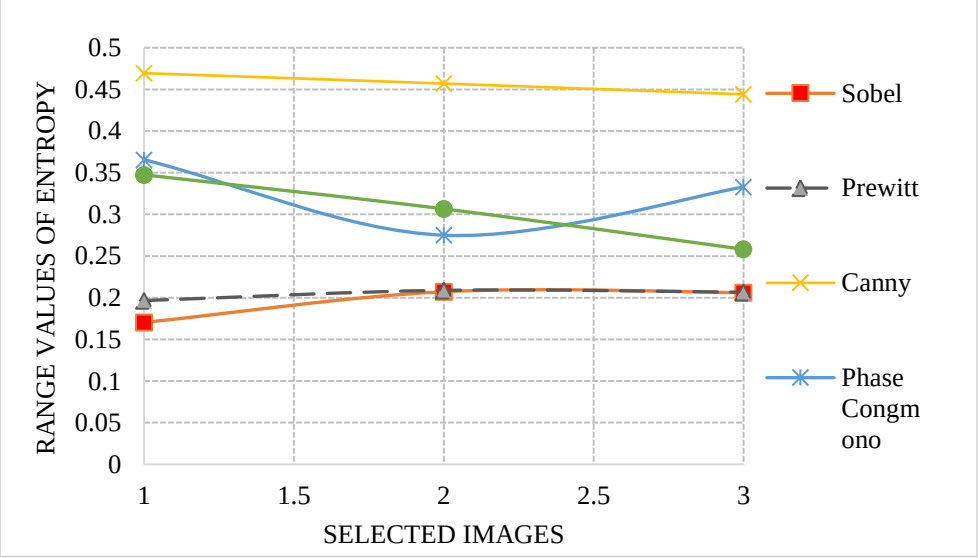


Fig. 5 The graphical analysis of result in comparison to previously proposed methods through entropy.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. CONCLUSIONS

The objective here is to acquire the effective and useful image information from the respective contours. Thus, it is essential to accurately detect object edges present in the image. To overcome the issues of edge connectivity from traditional approaches, a new edge detection algorithm is proposed in this paper. By the proposed algorithm following outcomes has been observed:

- Commendably discard of the inadequate evidence and only crucial edges are taken into consideration.
- Continuation in edges and thin edges has been achieved.
- Loss of crucial information has also been prevented.

REFERENCES

[1] R. Gonzalez and R. Woods, *Digital image processing*. 2002.
[2] B. Gardiner, S. A. Coleman, and B. W. Scotney, "Multiscale edge detection using a finite element framework for hexagonal pixel-based images," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1849–1861, 2016.
[3] S. Gupta and S. G. Mazumdar, "Sobel Edge Detection Algorithm," *Int. J. Comput. Sci. Manag. Res.*, vol. 2, no. 2, pp. 1578–1583, 2013.
[4] S. E. E.- Khamy, F. Ieee, and M. Lotfy, "A MODIFIED FUZZY SOBEL EDGE DETECTOR," *IEEE Seventeenth Natl. Radio Sci. Conf.*, pp. 1–9, 2000.
[5] K. Beant; G. Anil, "Comparative Study of Different Edge Detection Techniques," *Int. J. Eng. Sci. Technol.*, vol. 3, no. 3, pp. 1927–1935, 2011.
[6] S. Lakshmi, "A study of Edge Detection Techniques for Segmentation Computing Approaches," *IJCA Spec. Issue "Computer Aided Soft Comput. Tech. Imaging Biomed. Appl.*, pp. 35–41, 2010.
[7] O. Ghita and P. F. Whelan, "Computational approach for edge linking," *J. Electron. Imaging*, vol. 11, no. 4, p. 479, 2002.
[8] J. Canny, "A Computational Approach to Edge Detection," *Pattern Anal. Mach. Intell. IEEE Trans.*, no. 6, pp. 679–698, 1986.
[9] P. Kovesi, "Phase Congruency Detects Corners and Edges," *DICTA*, pp. 1–10, 2003.
[10] P. Kovesi, "Image Features From Phase Congruency," *Videre A J. Comput. Vis. Res. MIT Press*, vol. 1, no. 3, pp. 1–3, 1995.
[11] K. Peter, "Invariant Measures of Image Features from Phase Information," *Ph.D. Thesis*, pp. 1–180, 1996.
[12] W. Gao and X. Zhang, "An Improved Sobel Edge Detection," *3rd IEEE Int. Conf. Comput. Sci. Inf. Technol.*, pp. 67–71, 2010.
[13] C. Deng, W. Ma, and Y. Yin, "An edge detection approach of image fusion based on improved Sobel operator," *4th Int. Congr. Image Signal Process.*, vol. 3, no. 2, pp. 1189–1193, 2011.
[14] M. Cal and V. Di Mauro, "Performance Analysis of Roberts Edge Detection Using CUDA and OpenGL," pp. 55–62, 2016.
[15] L. Jiang, Y. Ye, and G. Xu, "Optik An efficient curve detection algorithm," *Opt. - Int. J. Light Electron Opt.*, vol. 127, no. 1, pp. 232–238, 2016.
[16] O. Prakash Verma, B. Neetu Agrawal, and B. Siddharth Sharma, "An Optimal Edge Detection Using Modified Artificial Bee Colony Algorithm," *Proc. Natl. Acad. Sci. India Sect. A Phys. Sci. Springer*, vol. 86, no. 2, pp. 157–168, 2016.
[17] S. Kavita and C. Vinay, "Computational intelligence in data mining—volume 1: Proceedings of the international conference on CIDM, 5-6 december 2015," *Adv. Intell. Syst. Comput.*, vol. 410, no. 1, pp. 159–169, 2016.
[18] K. S. and B. R. K. P. Hinduja and Abstract, "Edge Detection on an Image using Ant Colony Optimization," *Adv. Intell. Syst. Comput. Springer*, vol. 380, pp. 593–599, 2016.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Jaskaran Singh completed his undergraduate in Computer Science and Engineering from Guru Gobind Singh Indraprastha University, India. He is currently working as machine learning engineering in MediaAgility. His areas of interest include Digital Image Processing, Signal Processing, Machine Learning and artificial intelligence.

Bhavneet Kaur received the B.Sc. (H) degree in computer science from Delhi University, India and MCA degree from Sikkim Manipal University, India in 2014. She is currently working towards the Ph.D. degree in Computer Applications at Chandigarh University. Her research interests are digital image processing, computer vision and computer graphics. She is a lifetime member of ISTE, IEAE, IAENG, IASTER and ICSES.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fuzzy Expert System to diagnose Psoriasis Disease

Divya Mishra
Research Scholar
Uttarakhand Technical University
divya19aug@gmail.com

Dr.Nirvikar
Associate Professor
College of engineering roorkee
nirvikarlohan@yahoo.co.in

Dr.NeeluJyoti Ahuja
Sr.Associate Professor
University of Petroleum and energy studies
neelu@ddn.upes.ac.in

Deepak Painuli
Assistant Professor
Quantum school of technology
Deepak.painuli@gmail.com

ABSTRACT

Now a days skin disease is very common to all. There are various kinds of skin diseases that can affect one's life. The present research will be helpful in diagnosing the psoriasis which is a kind of skin disease. Fuzzy rule based system is used with Matlab to design rules and to diagnose the disease. Symptoms have been used to define the membership function.

Keywords: Rule based system, fuzzy logic, medical diagnosis, disease

I. Introduction

Artificial Intelligence (AI) is a growing field of computer science that is used to make such intelligent machines that perform the task normally performed by human experts.
Many applications of AI have been used to solve various complex problem, one of which is expert system [1]. Expert System is that system which stores expertise of domain experts in knowledge base and uses that expertise to solve user problem that previously solved by human experts. One of the crucial field in which Expert system is working is medical field [2].
Expert system is playing an important role in diagnosing diseases.A Fuzzy expert system is proposed to diagnose psoriasis disease.

II. Rule Based Fuzzy System

Fuzzy expert systems are developed using method of fuzzy logic which deal with uncertainty. The technique which uses the mathematical theory of fuzzy sets, simulates the process of normal human reasoning by allowing the computer to behave less precisely and logically than conventional computers. Fuzzy logic is fundamentally a multi-esteemed logic that enables median esteems to be characterized between regular assessments like yes/no, genuine/false, dark/white, and so on. Ideas like rather warm or entirely cool can be planned numerically and algorithmically prepared. Along these lines an endeavor is made to apply a more human-like state of mind in the programming of PCs ("soft" computing) [2].

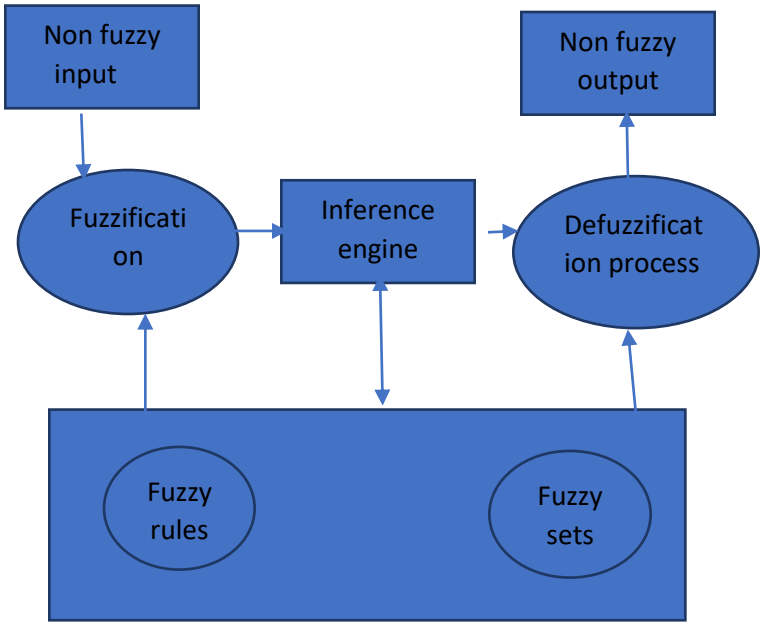


Figure 1

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. Psoriasis

Psoriasis is a sort of skin condition that expands the life cycle of skin cells. It makes cells develop quickly on the surface of the skin. The additional skin cells shape scales and red fixes that are irritated and in some cases painful. Psoriasis is a constant infection which regularly goes back and forth. The fundamental target of psoriasis treatment is to prevent the skin cells from developing so rapidly. There is no cure for this kind of skin disease, but symptoms can be managed. Lifestyle measures, for example, saturating, stopping smoking and overseeing pressure, may help.[3,19]

Psoriasis patches can extend from a couple of spots of dandruff-like scaling to real emissions that cover vast regions.

Most sorts of psoriasis encounter cycles, flaring for a large portion of a month or months, by then fading away for a period or not withstanding going into complete reduction. [3]

IV. Related Work

This paper highlights the challenges to diagnose psoriasis disease and provides the method that is used to identify the disease using fuzzy rule based system.

This paper identifies research gaps and provides a strong encouragement for further research in this field.It can also helpsmedical experts to identify disease as this paper focuses on the fuzzy values that helps in identifying the particular skin problem.

For this following survey is done which includes various methods suggested by the authors for diagnosing diseases.

Varinder[13] focused on diagnosing dengue fever with the help of fuzzy expert system. In this paper some lab features and clinical symptoms have been reported that tries to diagnose the disease.

Rangkuti et al.[18] focused on diagnosing the skin disease with the recommendation of medicines. The author uses fuzzy subtractive clustering algo to provide recommendation of treatment.

Stanley et al.[14] discusses about fuzzy logic based color histogram analysis technique for discriminating skin lesions from malignant melanomas in dermatology clinical images.

Akimoto et al.[15] discuss about fuzzy reasoning system to diagnose skin disease. The author defines the fuzzy membership functions using various texture-based features obtained from reference images. After that the author uses the classifier for disease identification.

Castellano et al.[12] focused on diagnosing dermatological diseases by neuro fuzzy system.The author uses neuro fuzzy system named,KERNEL.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Anbarzadeh et al.[16] discusses about web based fuzzy diagnosis system and evaluates five diseases with sort throat symptoms.

Jeddi et al.[17] proposes a diagnostic value of skin disease using expert system. The author diagnosis the skin disease like-pemphigus vulgaris, lichen planus, basal cell carcinoma, melanoma and scabies.

Saikia et al.[13] proposes an expert system to diagnose early dengue disease using fuzzy inference system. The author had taken the physical symptoms and medical tests reports as input variables and convert these into fuzzy membership functions.

In the above papers the authors have used various AI techniques for disease diagnosing like fuzzy logic, medical rule based system, neuro fuzzy system. In contrast this paper focuses on diagnosing psoriasis disease. A fuzzy rule based system is used to define the rules and based on those rules the MATLAB tool is used to define symptoms as membership function and find out the result based on those membership functions**.**

V. Implementation

Before implementation the collection of data is required regarding psoriasis disease. The dataset was prepared from the results of various symptoms of psoriasis disease. Following symptoms are used as input variables:

•Red thick spots on the skin [4]

•Small scaling spots usually found in youngsters[4]

•Dry, split skin that may deplete [4]

•Itching, soreness[4]

•Thickened, emptied or wrinkled nails[4]

•Expanded and firm joints [4]

The present work on Fuzzy based technique is performed in MATLAB R2016b. MATLAB also known as Matrix Laboratory, is used to model complex system by using logic rules and implementing these in fuzzy system. Fuzzy deduction is the way toward planning the mapping from an offered contribution to a yield utilizing fuzzy logic. The mapping is utilized to give a premise from which choices can be made. Figure 2 gives the depiction of MATLAB window while utilizing Fuzzy Inference System manager for 3 sources of info and 1 output.[7, 8 and 11].
Figure 2(a) ,(b) ,(c) ,(d) ,(e) shows the symptoms in the form of fuzzy sets that defines the membership function.
Figure 2(f) shows the result after implementing the rules based on table 1 and using the rules to define membership function.

Red patches	Itching	Thicked ridged nails	Result
0.5	0.5	1.0	Very high
0.75	0.25	0.75	Very high
0.25	0.75	0.25	Very low
0.5	1.0	0.75	High
0.25	0.75	0.5	Low
0.5	0.5	0.5	moderate

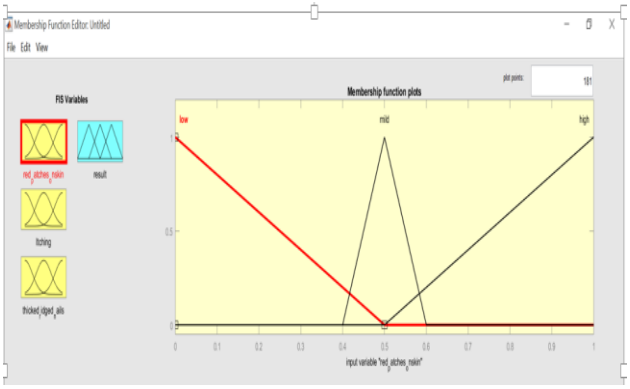
Table 1

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

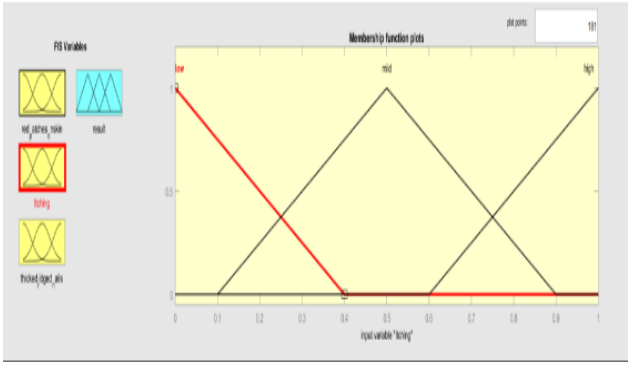
Values in table 1 have been referenced from file:///E:/phd/Fuzzy%20rulebased%20systems.html

Following are the rules designed on the basis of the data shown in table 1.

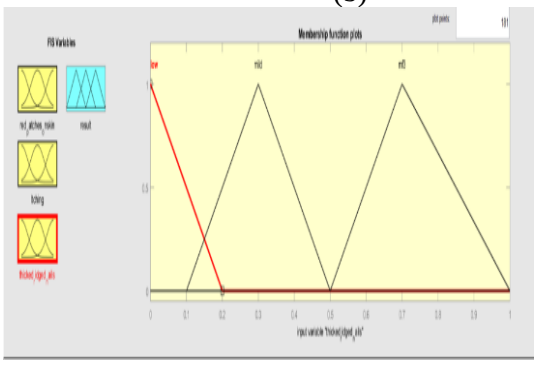
1. If red_patches on skin is low and itching is low and thicked_ridged_nails is low then psoriasis is low.
2. If red_patches is low and itching is low and thicked_ridged_nails is mild then psoriasis is low.
3. If red_patches is low and itching is mild and thicked_ridged_nails is low then psoriasis is ow.
4. If red_patches is low and itching is mild and thicked_ridged_nails is mild then psoriasis is mild.
5. If red_patches is low and itching is high and thicked_ridged_nails is mild then psoriasis is low.
6. If red_patches is low and itching is high and thicked_ridged_nails is high then psoriasis s mild.
7. If red_patches is mild and itching is low and thicked_ridged_nails is low then psoriasis is low.
8. If red_patches is mild and itching is low and thicked_ridged_nails is mild then psoriasis is mild.
9. If red_patches is mild and itching is mild and thicked_ridged_nails is mild then psoriasis is mild.
10. If red_patches is high and itching is low and thicked_ridged_nails is low then psoriasis is mild.
11. If red_patches is high and itching is low and thicked_ridged_nails is mild then psoriasis is mild.
12. If red_patches is high and itching is low and thicked_ridged_nails is high then psoriasis is high.
13. If red_patches is high and itching is mild and thicked_ridged_nails is low then psoriasis is mild.
14. If red_patches is high and itching is mild and thicked_ridged_nails is mild then psoriasis is high.
15. If red_patches is high and itching is high and thicked_ridged_nails is low then psoriasis is high.
16. If red_patches is high and itching is high and thicked_ridged_nails is mild then psoriasis is high.
17. If red_patches is high and itching is high and thicked_ridged_nails is high then psoriasis is high.



(a)

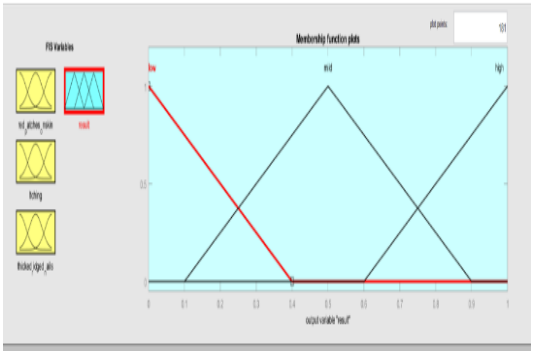


(b)

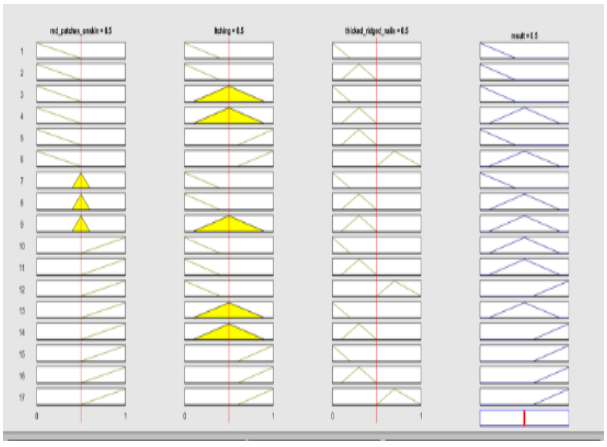


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

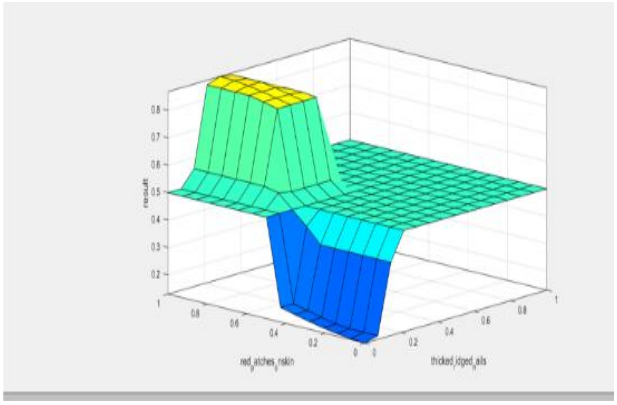
(c)



(d)



(e)



(f)

Figure 2

VI. Conclusion

This paper shows a fuzzy expert system in which conclusion of psoriasis sickness is displayed. Fuzzy rule based system is utilized to characterize the level of sickness. Information is spoken to by means of a formalism coordinating conventional guidelines and rules. This outcomes in better portrayal, since one can speak to more intricate relations amongst conditions, and encourages learning obtaining. All that a specialist needs to do is to decide the indications associated with diagnosing different ailments and the preparation sets.

VII. Future Scope

The defined work can further be proceeded for diagnosing of diseases other than psoriasis. The performed work gives correct prediction for skin disease. It can also be extended by using the neural network with the combination of fuzzy rule based system i.e; Hybrid system with the combination of both fuzzy and Artificial Neural Network (ANN).By using this approach the result will be more accurate.

VIII. References

[1]https://www.techopedia.com/definition/190/artificial-intelligence-ai

[2] Shu-Hsien Liao (2004), “Expert system methodologies and applications- a decade review from 1995 to 2004”,Elsevier

[3]https://www.medicinenet.com/psoriasis/article.htm#what_is_psoriasis

[4] G Ilczuk RMlynarski, A Wakulicz-Deja, A Drzewiecka, W Kargul. (2005), “Rough set techniques for medical diagnosis system,” Computers in Cardiology IEEE Conference on Control Applications Toronto, Canada

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[5] Ubeyli, ElifDerya (2010), "Automatic diagnosis of diabetes using adaptive neuro-fuzzy inference systems," Department of Electrical and Electronics Engineering, Faculty of Engineering, TOBB

[6] Yap, KeemSiah, Lim, Chee Peng&JunitaMohamad-Saleh. (2010), "An enhanced generalized adaptive resonance theory neural network and its application to medical pattern classification," Journal of Intelligent & Fuzzy Systems

[7] Baig, Faran, Khan, Saleem, Noor, Yasir, Imran. (2011), "Design model of fuzzy logic medical diagnosis control system," International Journal on Computer Science and Engineering (IJCSE).

[8] D. Bystrov, & J. Westin, "Practice, neuro-fuzzy logic systems, Matlab toolbox GUI", ch2, pp. 8-39.

[10] DarshanaSaika, Jiten Chandra Dutta (2016), "Early diagnosis of dengue disease using fuzzy inference system", IEEE

[11] www.mathworks.com

[12] G. Castellano, C. Castiello, A.M. Fanelli, C. Leone "diagnosis of dermatological diseases by a neuro-fuzzy system"

[13] VarinderPabbi (2015) "Fuzzy Expert System for Medical Diagnosis"

[14] R. Joe Stanley, Randy Hays Moss &Chetna Aggarwal "A fuzzy based histogram analysis technique for skin lesion discrimination in dermatology clinical images"

[15] Makio Akimoto ,Michio Miyazaki, Hee-Hyol Lee, Tomonori Nishimura, Mutsumi Tamura, &MichioMiyakawa (2009) " Using fuzzy reasoning to support a system of diagnosis of skin disease".

[16] SadafAnbarzadeh, HosseinDavari (2015) " Web based fuzzy diagnosis system and evaluation for five diseases with sort throat symptom".

[17] FatemehRangrazJeddiMasoudArabfard&Hamidreza Gilasi (2016) " The diagnostic value of skin disease diagnosis expert system"

[18] Abdul harisRangkuti, ZulfanyErlisaRasjid& Muhammad Iqbal Imaduddin (2015) " Face skin disease recognition using fuzzy subtractive clustering algorithm"

[19] https://www.webmd.com/skin-problems-and-treatments/psoriasis/default.html

[20]file:///E:/phd/Fuzzy%20rulebased%20systems.html

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Effect of Mobile device profiling in Mobile Computation Offloading

Mona ElKalamawy
Computer Science Department
Faculty of Computers and Information - Cairo University
Mona.elkalamawy@fci-cu.edu.eg

Abeer ElKorany
Computer Science Department
Faculty of Computers and Information - Cairo University
a.korani@fci-cu.edu.eg

Abstract— **although the great improvement and development in mobile manufacturing process, it's still considered as a resource limited device. Low battery capacity and weak processing power are constrains that meet mobile users incremental demands. Mobile Computation Offloading (MCO) is one of the available solutions to the current problem, by sending the computation-intensive applications to a powerful server which do the complex and resource intensive computations and send back the results to the mobile device. On the other hand, MCO may cause computation and communication overhead which lead to more power consumption. An important aspect of mobile offloading is to identify when offloading could take place. Currently, offloading decisions is mainly based on profiling performed on *individual devices*. In this work, a classification model is built to decide whether to offload or not based on the current device status. Crowed sensing data is utilized to learn the optimal offloading contexts/situations. The results show that CPU usage is the most influencing parameter that increases accuracy of offloading process by 86%.**

Keywords - Computation Offloading; Context-Awareness; energy efficiency; Crowed Sensing

I. INTRODUCTION

Mobile Computation Offloading (MCO) is an emerging paradigm that helps resource-constrained mobiles to run computation-intensive programs on mobile devices. That's done by offloading parts of program to be run on a server. Low battery capacity and weak processing power are constrains that meet mobile users incremental demands. Recently, researchers' attention is devoted to enhance the current situation and shrink the gap between user demands and mobile processing and power capabilities current problem, by sending the computation-intensive applications to a powerful server which do the complex and resource intensive computations and send back the results to the mobile device. MCO as a concept is similar to outsourcing in which the weak mobile device benefits from a powerful server. On the other hand, MCO may cause overhead which leads to more power consumption.

To explain how the offloading process goes on several offloading frameworks have been presented to focus on answering at least one of these Questions/perspectives:

1. What to offload? Methods/Classes that can be offloaded without affecting the application execution scenario are annotated. That annotation is done either manually by the developer or automatically through a partitioning module. The annotation is done only once at compilation time. Identifying "What to offload?" i.e. the modules in the application that could be offloaded is one of the concerns/ questions that computation offloading frameworks consider.
2. When to offload? Each time the program runs at the mobile device, the offloading framework analyze the current situation/environment's parameters to decide if it'll be beneficial to offload computation or run it locally. Decision engine is the module responsible for the offloading decision. Deciding "When to offload" is the second concern/question that should be investigated.
3. Where to offload? If a decision is taken to offload the code, another module should choose a server that will make real performance improvement. As mobiles have different types/capabilities, same server may be suitable for one device while being not beneficial for another. Depending on processing and memory configurations of both mobile and server. Choosing the suitable server to handle the mobile offloading request is the third aspect of the offloading process. Authors in [1][2] applied different techniques to find the most suitable server.

Though there is bunch of offloading frameworks available in academia that handle one or more of the above mentioned aspects, they lack the practical experimentations in real life mobile operating situations. Evaluation of these frameworks was done using limited number of mobile devices, which doesn't cover all the contexts/statuses that a mobile may work in.

To overcome this problem, crowed-sensing data is the solution. Crowed-sensing data is collected from thousands of mobile devices, so they cover a wide range of real life mobile operation contexts. By analyzing the collected data, we can distinguish the situations in which offloading will be beneficial from the situations in which offloading will cause overhead. By aggregating different execution contexts from

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

different devices with different connection types, we can learn the optimal realistic offloading situations. That's why crowed-sensing data is utilized in this work.

This work has two main contributions:

- Investigate crowed sensing data to learn the context features affecting the offloading decision, categorize them into three main categories and reason the effect of each feature on the battery consumption rate and offloading decision respectively.
- Develop a machine learning based decision engine that predicts the device statuses / contexts in which offloading will make an added value for battery consumption.

Therefore, the proposed framework for context aware computation offloading apply machine learning to select the appropriate mobile context parameters for offloading decision. To take a correct offloading decision, the decision engine should firstly scans the current status and then decide to offload or not. After gathering the required contextual information, the engine starts to analyze and apply machine learning model that utilize the current context parameters and predict weather the offloading decision will be beneficial.

The rest of the paper is organized as follows; In Section II we briefly give a literature review of computation offloading as a whole, and the current work on decision engines specifically. Then in Section III, we explain our proposed framework "the offloading decision engine" which mainly depends on identifying the most suitable mobile context parameters. In Section IV, we report the experiments using our proposed framework and the results. Finally, section V concludes the paper and outlines our future work.

II. Related Work

A. MCO Frameworks

During the current decade, more research focus was given to mobile computation offloading. Many offloading Frameworks have been developed to solve and cover one or more of offloading questions.

MAUI [4] is one of the pioneers frameworks. It proposes a strategy based on code annotations to determine which methods from a Class must be offloaded. The annotations are made by developers "Static Code Profiling". If suitable context is reached by profiler, it offloads the annotated piece of code. Due to static annotations, programs unable to adapt the execution of code in different devices. So MAUI's main focus is on what and when to offload.

Clone Cloud [5] tries to address static code profiling in MAUI by introducing code profiler that can choose pieces of bytecode of a given mobile component to run at remote server. Other parameters also influence when choosing a portion to code to offload, e.g. the serialization size, latency in the network, etc. So CloneCloud's main interest is to answer what & when to offload.

ThinkAir [6] tries to address MAUI's lack of scalability by creating virtual machines (VMs) of a complete smartphone system on the cloud. It provides an efficient way to perform on-demand resource allocation, and exploits parallelism by creating; resuming or destroying VMs. it focuses on the elasticity and scalability of the cloud and enhances the power of mobile cloud computing by parallelizing method execution using multiple virtual machine images. It uses static code annotation to mark offloadable parts.so we can see that thinkair focuses on answering "What, Where and when" to offload the code.

SmartPhone Energizer [7] Smartphone Energizer uses the benefit of supervised learning with a rich set of contextual information such as application, device, network, and user characteristics to optimize both the energy consumption and execution time of Smartphone's applications in a variety of contextual situations. The decision is taken so that offloading is guaranteed to optimize both the response time and energy consumption. The main objective of the framework is to decide "when to offload?"

Hermes [1] formulates an optimization problem from a given task dependency graph for the application, the objective of the formulation is to minimize the latency while meeting prescribed resource utilization constraints. It is of interest to find optimal assignments of tasks to local and remote devices that can take into account the application-specific profile, availability of computational resources, and link connectivity, and find a balance between energy consumption costs of mobile devices and latency for delay-sensitive applications. The framework Focus on answering the question where to offload.

ULOOF [8] a lightweight and efficient framework for mobile computation offloading. It is equipped with a decision engine that minimizes remote execution overhead, while not requiring any modification in the device's operating system. ULOOF main focus is what & when to offload.

MobiCOP [9] is an MCO framework that focus on enhancing reliability and scalability so it offers compatibility with most android devices available today. The decision engine here is based on a code profiler that's responsible for keeping track of past task executions and predicting the running time of future tasks.

Recent approaches utilize crowed-sensing data to help enhance decision accuracy. Crowd-sensing means collecting mobile device context samples from a large community of devices, this data is used to learn the optimal execution contexts that are most appropriate for offloading Carat [3] is an example of crowed-sensing datasets as it aggregate different mobile status parameters from a large community of users and report the energy drain for each device context.

EMCO [10] is a pioneer MCO framework that use crowed sensing approach to improve the prediction of the offloading decision. Instead of limiting to profiling individual devices, crowdsensing enables characterizing execution contexts across a community of users, providing better generalization and coverage of contexts. EMCO focus on answering what, when and where to offload.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Decision Engines in MCO Frameworks

Literature [11][12][13] shows different approaches that have been applied for offloading decision. In this subsection, we will highlight three of the known offloading decision approaches.

1. *Stochastic Process*: in which the mobile application is abstracted and modeled statistically; this help to predict and evaluate optimal execution conditions. Statistical methods like Markov processes, the Poisson process, and queuing theory. Authors in [14] proposed a semi-markovian Decision process (SMDP) to model their application and find the optimal offloading decision. While stochastic models can find optimal decision, they consume a non- neglected processing over-head.
2. *Heuristic-Based decision Engine*: as its name can explain, in this method the decision engine applies heuristic rules on the current application situation to decide whether it's beneficial to offload the application or not. Heuristics may be like: if two tasks are sharing the same logic and input parameters, they probably will take the same time to be completed. MObiCOP [9] used heuristics to build their decision engine by predicting the execution time and energy consumption of the offloadable module. MobiCOP offers performance improvements of up to 17x and increased battery efficiency of up to 25x
3. *Machine Learning*: the idea of machine learning is to learn from the history in order to predict the future. Thus, by keeping records of past execution contexts and their energy consumption, machine learning could be applied to predict energy consumption for new execution contexts as in [7]. Furthermore, machine learning was also applied to predict the offloading decision as in [15] the authors developed a decision tree classification model to predict the offloading decision given the current device context.

The proposed framework uses method three "Machine Learning" approach to identify the context features affecting the offloading decision as well as detect the most affecting feature.

III. Proposed Framework For Identification of Offloading parameters

Battery consumption rate is a key factor in the offloading decision; situations that consume high battery levels are more probable to be offloaded. Thus, different context parameters that highly affect the mobile battery consumption/drain rate should be considered. For example, high screen brightness increases the consumption rate keeping the other parameters constant. Different context parameter value affects battery drain in a different way. In this work, the impact of changes of each parameter is studied as well as its influence on the battery drain rate. The following sub-sections describe the details of each of those parameters and how the proposed framework utilizes them in the prediction process.

A. Extracting Parameters for offloading process .

There are many parameters that affect the offloading decision; which could be categorized into three categories:

- Mobile device parameters
- Environment parameters
- Application parameters

Each of those categories plays an important role in the offloading decision process; to make an accurate decision, all the above parameters should be studied. The more the engine knows the proper the decision. In the following subsections, details of those categories will be described.

1) Mobile device profile:

These are the parameters that describe the status of mobile components like:

- Screen brightness.
- CPU usage.
- Battery temperature.
- Battery voltage (capacity of battery in volts).

Each of these parameters affects the process of offloading; CPU usage can indicate the computation complexity of the running application; the higher the usage the more complex the computations and accordingly the higher the battery drain. Same could be applied for the screen brightness.
Moreover, Batteries tend to operate properly in normal room temperature (20-30 C), if the temperature is lower, the resistance inside the battery circuit increase which implies higher battery drain. If the temperature is higher, the performance of the battery decreases.

2) Environment parameters

That kind of parameters describes the environment that the device is operating in:

- Network type (Wi-Fi or Mobile Data).
- Wi-Fi signal strength.
- Wi-Fi link speed.
- Mobile data status (on / off).
- Mobile data activity (in, out, in-out, none).
- Mobile network type (LTE, GPRS …).
- Distance travelled (if the user is moving).

Each of above parameters affects the battery drain rate, and as a result, they impact the process of offloading. The network connection type differs in power consumption; mobile data tend to consume more battery power than Wi-Fi. Signal strength and link speed also impact the battery drain level. Weak signals force the mobile to use more power to send/receive data. Different mobile data network types (3G, 4G…) consume different energy levels. So all of these informative parameters enrich the model with the current mobile status, which help determining whether the energy drain level is normal or high. High energy drain contexts are favorable to be offloaded.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

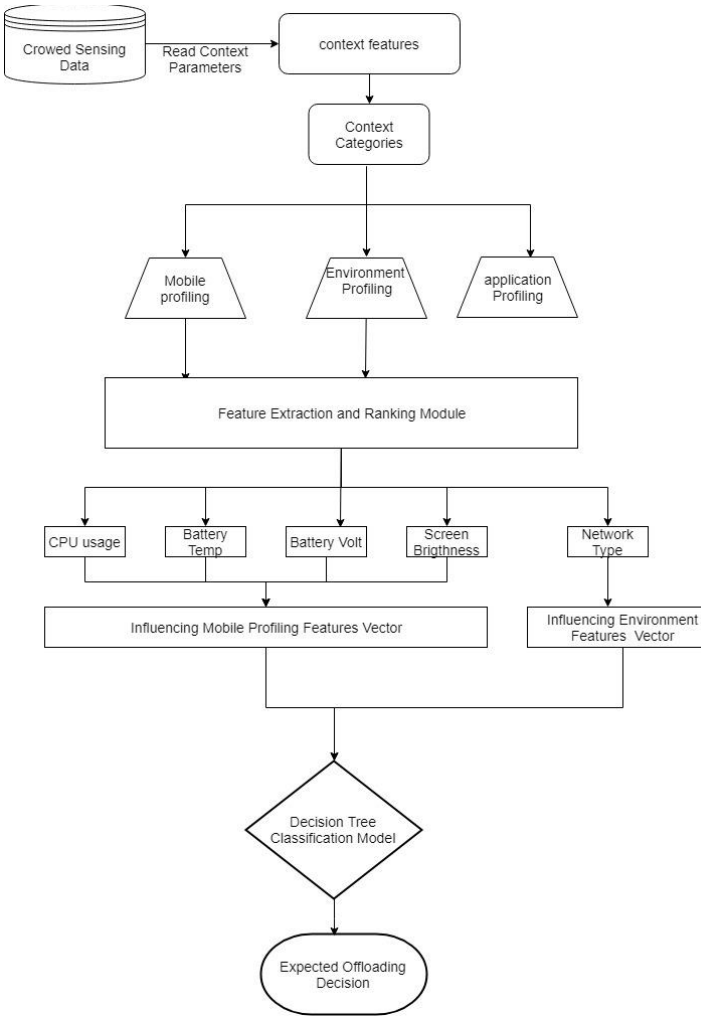


Figure 1 The Framework of For Identification of Offloading parameters

3) Application parameters:

These are the features that describe the application/software to be offloaded such as:

- Input size.
- No of instructions.
- Cohesion & coupling degree between modules.
- Expected execution time.

However, in this work we only focus on the first two categories of parameters, mobile and environment.

B. offloading prediction

To make the optimal offloading decision, we build a classification model to predict whether to offload or not, given the current context parameter values. The classifier main objective is to decide if the current status consume high energy rate or not. If the expected energy drain rate is high, the classifier decides to offload the computation. To build the classification model, a subset of the features is selected to train the classifier. Feature selection is a process in which a subset of the features is selected to train the classifier. This feature subset is the most relevant to the classification decision.

The objective of feature selection techniques is to simplify machine learning models in order to be easily understandable by researchers and at same time lessen the complexity, shorten the required training time and reducing over fitting of the learning model.

Decision tree [16] classifier is used to build the prediction model. It builds the classification model in a form like a tree structure, it divides the dataset into smaller subsets, and meanwhile it builds an associated decision tree incrementally. It has two kinds of nodes; leaf nodes and decision nodes. Leaf nodes represents the final classification decision (offload or don't offload), while decision node has branches that represent possible decision alternatives. Fig. 2 shows a part of the classification model decision tree

C. Feature Ranking

Features affect battery drain in different rates, in this section we investigate which features are the most influencing in the energy consumption and offloading as well. Mobile device profile parameters have a greater effect on the battery drain rates compared to environment parameters. CPU usage specifically has the greatest impact, which makes sense. Classification accuracy results decreased by 20% when training the classifier without CPU usage attribute as we will show in the next section. The second influencing features are battery temperature and voltage, as high or low temperatures are not favorable. Screen brightness comes at last position of influencing features' ranking. Environment features didn't affect the battery drain rate that much; network type (WiFi or mobile data) has the greatest impact.

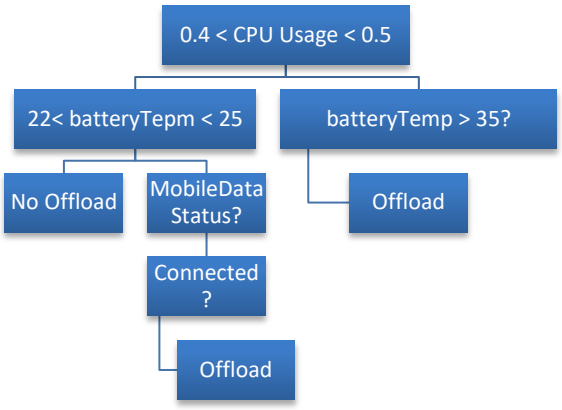


Figure 2: Part of Classification Decision Tree

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. EXPERIMENTS AND RESULTS

A. The Dataset "Carat"

Carat is a mobile application that tells the user what is using up the battery of the mobile device, whether that's normal, and recommends the user what he can do about it. After running Carat for about a week, user will start to receive personalized recommendations for improving the battery life.

Meanwhile, carat collects data from the community of users and save them in a dataset to figure out what are the applications / situations that consume the most energy.

Carat context factor dataset contains about 11 million records of different mobile contexts; each instance reports the current context feature values and the corresponding energy drain rate. Attributes are: CPU Usage, Battery Temperature, Battery Voltage, Screen Brightness, Network Type, Mobile Data Status, Mobile Network Type, WIFI Signal Strength, Distance Traveled, Mobile Data Activity, WIFI Link Speed, Battery Health and Roaming Enabled.
A subset of the features (CPU Usage, Battery Temperature, Battery Voltage, Screen Brightness, Network Type and Mobile Data Status) was used to train our classifier to fasten the training time. Selection of the subset enhanced the classifier accuracy results by 0.1% which indicates that some features may have a conflicting impact on the classifier.

B. Preprocessing

Carat dataset doesn't provide information about offloading decision, each instance of the data reports for a given context parameter values; the energy drain rate is X per second. In order to be able to utilize this dataset for the research purpose, an energy drain threshold should be estimated. Based on EMCO [10], where authors concluded that if the current context consume energy drain rate > 0.004 per second, then offloading will be favorable.

Furthermore, the dataset has noisy values make sense, such as CPU usage parameter with negative value. Thus, some meaningless records have been deleted. To ensure an accurate classification model, a sample of 60000 valid records was selected to train the classifier.

Selection Criteria:

- Screen brightness range [0 - 255].
- CPU usage range [0 - 1].
- Battery temperature [0 - 50].
- Battery voltage [0 - 5].

Discretizing Numerical Attributes

It may be easier for some machine learning algorithms to deal with discrete-valued / nominal attributes. This is done by choosing split points in continuous / numerical attribute values range and grouping all values in the same range to have one nominal value. Weka supports discretizing attributes through providing a discretize filter.

In addition, Weka has a built in classifier called "filtered classifier" that does these two steps in sequence. It first filters the training data with the desired filter, then it feed these filtered data into a classifier, it also applies the same filtering technique among test data.

C. Classification Model Sequence

- Dataset is preprocessed to select the experimentation sample with no noisy values.
- Records are classified, if a record has energy drain rate greater than 0.004 then it is marked "Yes", else it's marked "No".
- Train and test sets are balanced; 50% of records are of "yes" category and the other 50% are of "No" category.
- Discretization of numerical attributes to be nominal is done.
- Weka [17] machine learning workbench is used to build our classification model.

D. Classification Experiments

1. Base line model was developed to estimate preliminary classification accuracy results. A random sample of carat data with no preprocessing was examined with different classifiers (MLP is MultiLayer Perceptron & SVM is Support Vector Machine). Figure 3 shows the accuracy and f-measure results of the experimented classifiers without data cleaning. Results showed that decision tree classifier achieved the most accurate results with respect to the others
2. Preprocessing the data helped a lot to achieve higher classification accuracy, as shown in figure 4, when a preprocessed sample is tested among same classifiers, the accuracy results enhanced by about 10%. Furthermore, 5 fold validation was applied to ensure the accuracy of results. Accuracy and f-measure are almost the same for most of the classifiers; this is because the sample test and train data were balanced for each category. Decision tree classifier still achieving the highest accuracy.
3. From the above experiments, we concluded that decision tree classifier is the most accurate one. Consequently, step 3 was to discretize the numerical attributes before building the classifier. The accuracy results jumped to **86%** when discretizing technique was adopted.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

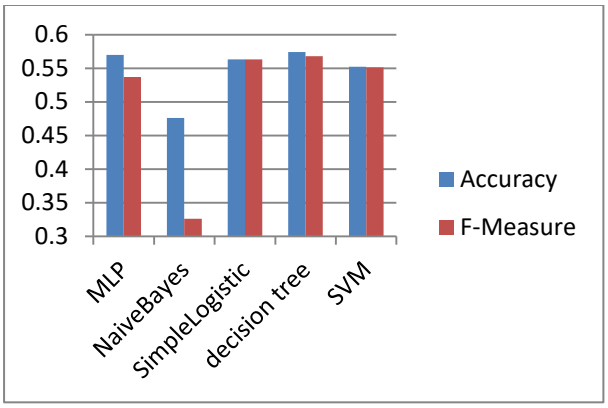


Figure 3 Base Line Model (without data cleaning)

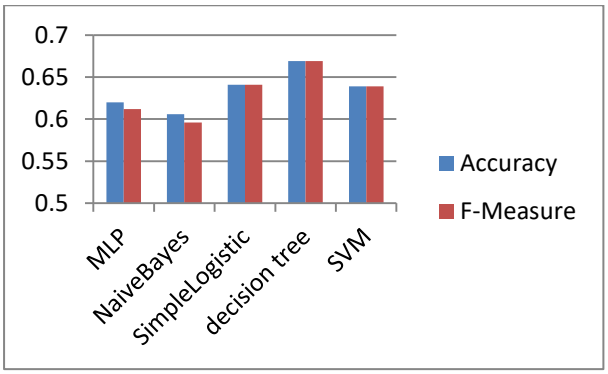


Figure 4: Comparison between different classifiers

E. Experminting The Influencing Parameters:

Experiments showed that mobile profile features have a greater impact on the classification decision accuracy than the environment features. CPU usage attribute has the greatest impact. That's make sense, as high CPU usage may indicate the excessive processing requirements of the running applications which implies that offloading the computations is a favorable decision. The second important feature is the battery temperature, as we explained above, batteries normal operation temperature is between 20-30 C. higher or lower temperatures affect the battery drain rate badly.

Feature Selection is done to prove the importance of each feature as well as to detect the most decision-influencing context parameters. Experiments showed the most impacting features are (in order):

- CPU Usage
- Battery Temperature.
- Battery Voltage.
- Screen Brightness
- Network Type.

When using the above features to train our decision tree, accuracy results enhanced by 0.1 %. Fig. 5 shows how each of the parameters affects the decision accuracy. CPU usage has the greatest impact as expected.

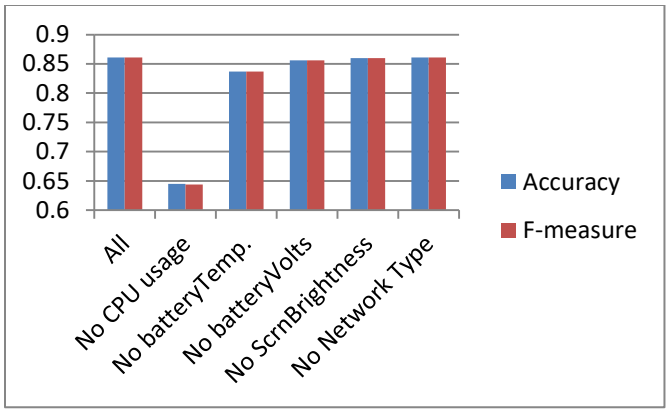


Figure 5: Offloading Influencing Parameters

V. Conclusion and Future Work

Because battery consumption rate is a key factor that affect the offloading decision. This work has two main contributions; the first is that we investigated the different execution parameters that impact the battery drain rate and consequently the offloading decision. The second contribution is that we developed a machine learning model based on decision tree classifier to predict whether to offload or not based on the current context parameters. The classifier can predict make the decision with accuracy 86.1%.

Context parameters are categorized into three categories; mobile device profile parameters, environment parameters and application parameters. Mobile device parameters have the strongest impact on the battery drain rate; CPU usage specifically is the most influencing parameter then battery temperature.

The proposed framework for optimizing offloading parameters focused on the first two types of execution context parameters (mobile and environment). So it will be further improved in the future by investigating the third kind of context parameters; the application parameters.

VI. References

[1] Y. H. Kao, B. Krishnamachari, M. R. Ra, and F. Bai, "HERMES: Latency Optimal Task Assignment for Resource-constrained Mobile Computing," *IEEE Trans. Mob. Comput.*, vol. 16, no. 11, pp. 3056–3069, 2017.

[2] X. Chen, S. Chen, X. Zeng, X. Zheng, Y. Zhang, and C. Rong, "Framework for context-aware computation offloading in mobile cloud computing," *J. Cloud Comput.*, vol. 6, no. 1, pp. 1–17, 2017.

[3] A. J. Oliner, A. P. Iyer, I. Stoica, E. Lagerspetz, and S. Tarkoma, "Carat: Collaborative Energy Diagnosis for Mobile Devices," *Proc. 11th ACM Conf. Embed. Networked Sens. Syst.*, p. 10:1--10:14, 2013.

[4] E. Cuervo, A. Balasubramanian, and D. Cho, "MAUI : Making Smartphones Last Longer with Code

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Offload," *Proc. 8th …*, vol. 17, pp. 49–62, 2010.

[5] B. B. Chun, S. Ihm, P. Maniatis, and A. Patti, "CloneCloud : Elastic Execution between Mobile Device and Cloud," *Proc. sixth …*, pp. 301–314, 2011.

[6] S. Kosta, A. Aucinas, and R. Mortier, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," *2012 Proc. IEEE INFOCOM*, pp. 945–953, 2012.

[7] A. Khairy, H. H. Ammar, and R. Bahgat, "Smartphone Energizer: Extending Smartphone's battery life with smart offloading," *2013 9th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2013*, pp. 329–336, 2013.

[8] J. L. D. Neto, S. young Yu, D. F. Macedo, J. M. S. Nogueira, R. Langar, and S. Secci, "ULOOF: a User Level Online Offloading Framework for Mobile Edge Computing," *IEEE Trans. Mob. Comput.*, 2018.

[9] J. I. Benedetto, G. Valenzuela, P. Sanabria, A. Neyem, J. Navón, and C. Poellabauer, "MobiCOP: A Scalable and Reliable Mobile Code Offloading Solution," *Wirel. Commun. Mob. Comput.*, vol. 2018, 2018.

[10] H. Flores *et al.*, "Evidence-aware Mobile Computational Offloading," *IEEE Trans. Mob. Comput.*, vol. XX, no. XX, 2017.

[11] B. Zhou and R. Buyya, "Augmentation Techniques for Mobile Cloud Computing," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 1–38, 2018.

[12] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," pp. 1–37, 2017.

[13] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," 2017.

[14] S. Chen, Y. Wang, and M. Pedram, "A semi-Markovian decision process based control method for offloading tasks from mobile devices to the cloud," *GLOBECOM - IEEE Glob. Telecommun. Conf.*, pp. 2885–2890, 2013.

[15] S. M. A. Karim and J. J. Prevost, "A machine learning based approach to mobile cloud offloading," *2017 Comput. Conf.*, no. July, pp. 675–680, 2017.

[16] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software : An update The WEKA Data Mining Software : An Update," no. November, 2008.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Smart Water Management in Smart Cities based on Wireless Sensor Network (A survey)

Tirtharaj Sapkota
Dept. of Computer Science & IT
Assam Don Bosco University
Guwahati, India
tirtharaj.sapkota@purbuniv.edu.np

Bobby Sharma
Dept. of Computer Science & IT
Assam Don Bosco University
Guwahati, India
bobby.sharma@dbuniversity.ac.in

***Abstract*— Smart water is a key building block of smart city. Smart Water Management is a technology-driven water management system. It refers to optimization of water distribution and waste water network in order to reduce the operational and energy cost. It emphasizes that water and the energy used to transport it should be managed effectively. A report from World Health Organization (WHO) argues that a large section of the World's population lives in cities and is estimated that 70% of the world's total population will live in cities by 2050. Eventually water consumption in cities will increase. Therefore it is very important that water distribution and waste water-recycling should be done in an optimized manner in the cities around the world. Wireless Sensor Network (WSN) is a great enabler for this purpose. Water loss management is becoming increasingly important as supplies are stressed by population growth or water scarcity. WSN is a technology which can be adapted to speed-up the water-cycle and reduce the water loss. It can be used in real-time monitoring and control of such water distribution system. This paper reviews the use of WSN in Smart Water Management and addresses related problems faced by some of the water-smart cities and points out future opportunities that these cities have.**

Keywords— Smart Water Management, Waste water Network, Wireless sensor Network, Water-smart City, Smart Water meter, Energy Management.

I. INTRODUCTION

Water is a natural resource which is vital for life and environment. Because of its scarcity, managing the water has become crucial to meet the water demand of each society. Inadequate water quantity or bad water quality can be a limiting factor in economic development of any country. At present global urbanization is increasing at an exponential rate. A report from World Health Organization (WHO) argues that "around 70% of the world's total population is expected to live in cities by 2050 "[3]. So it is very important for the cities to become smart. For a city to become smarter, in general, means to increase the efficiency of the infrastructures and services to a level that was never previously achieved. The cities will have to understand the power of data and the importance of technology to optimize the resource consumption and reduce the amount of waste by both the producers and consumers of goods and services [1]. To become a smart water city, the cities will have to make use of water sensors, which eventually will form a wireless sensor network (WSN). WSN is a new attempt to maintain and control public drinking water resources and best to meet the increase demand for pro-active water quality management because WSN provide maximum return on their technological investment [10].

The water sensors in WSN can communicate among themselves to measure flow rates at different points in water network to detect leakages, as well as to monitor water quality issues including pH, dissolved oxygen and turbidity in the network. Smart water meters can also be used that allow the consumers to see their water consumption is real time and compare it with their neighbor's level of consumption. A Smart Water Meter is a normal water meter linked to a device that allows continuous electronic reading and display of the water consumption. It negates the need to manually read the meter dial [2]. These meters can also transmit appropriate data that will help in billing. These smart meters can also alert the consumers, with precise reason, if their consumption is very high [1].

Smart Water Management is a technology-driven water management system. The use of Information and Communication Technology (ICT) and real-time data processing system are indispensible. The potential application of smart water management system is wide and includes solutions for water quality, water quantity, efficient irrigation, leaks, pressure and flow, floods, droughts and much more [4].

Wireless Sensor Network can be great enabler for Smart Water Management. WSN generally have large number of sensors that are placed very close to each other [5].They are place to each other because of their limited bandwidth otherwise communication among them will not be possible. Typically WSN contains hundreds and thousands of sensor nodes and these sensors have the capability to communicate either among each other or directly to a base station (BS). A greater number of sensors allows for sensing over larger geographical regions with greater accuracy. Each sensor node comprises sensing unit, processing unit, communicating unit and power unit. Sensor nodes coordinate among themselves to produce high quality information about the physical environment. Wireless sensor networks are a viable solution for monitoring the condition, in particular the pressure and hence leaks, of buried water pipelines. They have advantages over other methods. One of the advantages is that they do not need human intervention for monitoring faults. Also failure of individual nodes does not hamper the working of the entire network.

II. LITERATURE REVIEW

A sensor node is composed of the sensing element, signal amplification, and filtering system with specific purpose software for data manipulation. Wireless sensors

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

have a special wireless transmission element. General water quality parameters such as pH, chlorine, temperature, flow, and turbidity are commonly monitored using wireless sensors. Primary sensing component is one of the main elements of water quality detection sensor, having physical, chemical and biological characteristics of a variety of different materials [6].

Lima a desert city experiences only 1cm of rainfall every year which is very low as compared to their population which is around 8 million[7]. The people of Lima mostly depend on glacial melt water and it has been reported that the glacier are shrinking and they will vanish completely in the next 40 years. In 2007, the government officials of Lima decided to take some action and to cope with these severe water shortages. The ministry of housing along with World Bank's water and sanitation program decided to promote water saving by educating people and by providing environmental awareness programs. They also decided to use latest information and communication technologies to cope with these problems.

In [7], water management in Cape Town was highlighted. Cape Town is internationally known for its efforts in water management. The city was able to decrease the water consumption by 30% in the last 15 years during that period population growth was over 30%. The city was able to improve its water management process by advocating two principle concepts. i) Convincing people to use less water. ii) Using Information and communication technologies to use water properly. Adjusting the water pressure, putting new pipes, detecting and repairing leakage were some of the activities carried out to improve the water management. In 2011, the city was able to change more than 20 thousand faulty water meters and was able to provide intensive training on water conservation to more than 60 schools.

In [8], smart water management around the North America region has been focused. The smart water management in North America is expected to grow from $1.77 billion in 2013 to $3.64 billion in 2018. The water management solutions with ICT help in better management of authorities' assets and distribution network reduce repairing cost. The Government in North America has developed strong rules and regulations for water management in consent with environmental standards. North America region has many ongoing smart city projects and smart water management is the heart of all these projects.

The US and Canada have already started large scale project on water management and distribution system to meet the rising demand of water resources [8].

In [9], the authors have considered wireless sensor network as an effective option of the conventional water monitoring system. Wireless sensor networks are relatively affordable and allow measurement to be taken from remotely in real-time and with minimal human intervention. The traditional method of water monitoring were laborious, time consuming and lacks real-time result to promote proactive response to water contamination.

In [11], reduction in water consumption in England through the use of wireless sensor network has been discussed. The use of WSN help to optimized the water facility by noticing the leaks or by examining the distribution of water in the network which help people to make informed decision about how water is managed efficiently. For example, the sensors in the WSN can find leaks in the water pipe and instantly inform engineers or the responsible authorities to take specific action. This is much valuable considering the fact about 3.3 billion liter of water is wasted each day in England and Wales due to poor infrastructure.

Smart water technology is helping consumers to check their water consumption level remotely through various applications. It informs the consumer when there is a leak, allowing people to find solutions much earlier and prevent water from being wasted unnecessarily. Smart water tools are being more accepted now a days, due to which consumer can control water consumption and make decisions based on the data they are seeing, helping them to use water more sustainably as well as cutting costs [11].

III. Benefits of Wireless Sensor Network over Conventional Method for Smart Water Management

A wireless Sensor Network has several benefits over traditional methods [10]. Some of them are:

1) Infrastructure: A WSN can be very easily applied in water bodies like water tank, rivers or ponds. Even in dynamic environment under the water, WSN has the ability to monitor various activities.
2) Delay Tolerance: Application of smart water management like Water quality monitoring is a real-time activity where delayed responses are of no use. So WSN can be applied to handle such problems to get responses immediately whenever an event occurs.
3) Availability of various sensors: Unlike traditional system, where precision limit vary from one instrument to the other, in case of WSN we can very accurately measure the various constituent of water.
4) Human intervention is not required: WSN consist of intelligent sensors which can operate without human intervention and so they are very suitable for smart water management.
5) Real time monitoring: WSN is perfectly suitable for monitoring and control of real time aspect of quality deterioration.
6) Possibility of alerts or early warning system: The WSN overcomes the problem of immediate alert or early warning system of conventional manual

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

monitoring. Lack of early warning in water management can be unsafe and may be life threatening in some cases.

7) Installation and maintenance: WSN have the capacity of self-organizing or reorganizing themselves without human intervention. This can greatly reduce the cost involved in installation and maintenance of various traditional devices.

IV. CHALLENGES OF IMPLEMENTING SMART WATER MANAGEMENT

There are many challenges of implementing Smart Water Management which were observed in research done so far [12][13]. They are summarized as follows:

1) High cost of installation, implementation and maintenance of Smart Water Management system.
2) Unavailability of hardware and software platforms for real-time data collection, processing and monitoring.
3) Lack of software program and algorithms that detect complications and pollution when performing water quality testing.
4) Lack of energy efficient techniques that allow the sensor network to run for a longer period of time to increase the reliability of smart water systems.
5) Lack of standard policy and regulation for smart water management.
6) Lack of national and international standards and reference architectures.
7) Lack of technology architecture for tasks like systems integration, communication and event handling.

V. CONCLUSION

Most cities depend on few sources of water such as river and ponds. Water resources around the cities in the world are diminishing due to increased water consumption as population in urban areas is increasing. There is no water metering in most of the cities. As a result the percentage of water loss is very high and there is no any means to track this loss. In order to improve water management and prevent loss, cities need a technology driven integrated approach to meet the water requirement of their citizens. Smart water management is a technology driven solution for managing water in cities around the world. Using technology like Wireless Sensor Networks, monitoring of water facility can be performed using real time data. Pipe line monitoring, pump station monitoring, water quality test are few activities that can be accomplished using wireless sensor Network. Smart water meter along with WSN encourages consumer to actively take part in water management and prevent water loss. In this paper we have focused on few smart cities around the world that use smart water management to overcome their water problems like shortage of water , loss of water due unavailability or poor infrastructure. Like the cities discussed in the paper, other cities should immediately start using the technology driven methods in the development of rules and regulation for the growth of the water use and recycling of water and awareness of the water resource. This will help the cities in successful implementation of a smart water management and eventually become a smart city.

REFERENCES

[1] http://markandfocus.com/2017/01/25/smart-city-smarter-water-management.

[2] G. Hauber-Davidson, E Idris, "Smart Water Metering", Water, May 2006, pp.38-41.

[3] http://waterworld.com/articles/print/volume-29/issue-12/water-utility-management/smart-water-a-key-building-block-of-the-smart-city-of-the-future.html.

[4] http://www.iwra.org/swm.

[5] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, A survey on Sensor Networks, IEEE Communications Megazine, August2003, pp 102-114

[6] J. Dong ,G.Wang, H. Yan, J. Xu, X. Zhang, "A survey of smart water quality monitoring system", Springer, 6th jan 2015.

[7] Five of the best water-smart cities in the developing world, The Guardain (international Edition), Feb 2016.

[8] http://www.micromarketmonitor.com/market/north-america-smart-water-management-1271665222.html, retrived on 31st August,2018.

[9] M. Pule, A. Yahya, J. Chuma, "Wireless sensor networks: A survey on monitoring water quality", Journal of Applied Research and Technology,Volume 15, Issue 6, December 2017, Pages 562-570.

[10] S. Verma, P. Choudhary, "Wireless Sensor Network application for water quality monitoring in India", 2012 National Conference on Computing and Communication Systems (NCCCS), 2012.

[11] http://www.hitachi.eu/en/social-innovation-tories/communities/smart-water-smart-cities, retirived on 1st September 2018.

[12] B.Gebremedhin, "Smart Water Measurements: Literature Review", Water-M Project Center for Wireless Communications (CWC), June 13, 2015.

[13] H. Jan Top. Smart grids and smart water metering in the Netherlands, 2010.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A novel modernization approach for migration to native cloud application: A case study

Kh. Sabiri, F. Benabbou, M.A. Hanine, A. Khammal and Kh.Akodadi

***Abstract*—A number of enterprises are passing to cloud computing is in growth, due to their many advantages, such as business agility, flexibility, cost reduction, scalability and easy management of resources. Otherwise, it may be not enough of the migrated application to benefit from cloud benefits, since the legacy application is just bundled with a virtual machine image or a container. In fact, this migrated application tends to be considered as a native cloud application without respecting of all their properties. It becomes a must to change the architecture of the legacy application, in order to be more agile and flexible, before deploying into the cloud. In this work, we propose a modernization iterative and incremental process to overcome the listed issues above. This process is based on the concept of smart use case combining with the architecture driven modernization (ADM) approach. In order to test approved approach, we give a case study and explain the process carried out in each phase.**

***Index Terms*—Enter key words or phrases in alphabetical order, separated by commas.**

I. Introduction

Cloud computing offers to enterprises plenty of advantages, such as scalability, elasticity, business agility, reliability and flexibility [1]. In order to take its advantage, there is a need to adopt a modernization approach that allows to benefit from it. Yet many companies, in their migration to the cloud, have only been able to take advantage of resource scalability [2] [3] following a pay as you go model [4] [5]. This type of migration revolves around the concept of virtualization, as applications are packaged with an image or a container of a virtual machine. These applications tend to be considered as native cloud applications, without respecting its properties [6] as Isolation of state, Distribution, Elasticity, Automated Management and Loose coupling. They focus only on improving the old application technologies, to take one side advantage of the cloud paradigm and its technological evolution. This latter helps to lower their IT costs only, without interest, in the modernization aspect of the herald applications, in order to make it agile, flexible, fast and so on. This aspect of modernization can ensure business continuity requirements as well as the longevity of IT.

Indeed, the migrated application retains the same complexity as before. All functionality of the application is grouped into a single unit, while it can be developed in a modular way. It is also necessary to redeploy the application entirety for any case of updates, maintenance or addition of new business requirements. Also, if a single crash occurred, it will strike all the application services. Additionally, the scalability of the application would not be possible without a scalable architecture.

Most current approaches for cloud migration focus on automated migration by applying model-driven approaches [7] [8] and knowledge reuse through migration models [9]. [10] [11]. They didn't focus on the native cloud application architectures and its agility, but many studies [9] [12] [13] concluded that virtualization alone is not enough to take full advantage of the cloud computing paradigm.

In this context, and in view of the similarity of our proposed approach, we have made a detailed comparative study of these approaches by identifying the weaknesses and strengths of each approach [14]. Despite the fact that the methods presented are based on the model paradigm, the resulting application, migrated and deployed in the cloud, does not conform to the native cloud application architecture. They have particularly focused on how to adapt and adjust the legacy application to be deployed in the cloud environment.

In this paper, we propose an agile approach to modernize and to migrate a legacy application to the cloud environment. This new agile modernization approach supported by model-driven approaches, which place the models at the heart of the software engineering process. In addition, our approach facilitates convergence towards the goal of including all the benefits promised by the cloud environment, making them effectively exploited by the user, such as flexibility, business agility, scalability, elasticity and so on.

In the first place, we will define the basic concepts on which our approach has been based, in particular, Domain Driven Design, use cases. Second, we detail the new modernization process and its different phases. In the last, we will present the application, which is going to be the subject of

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This **work** was supported in part by the U.S. Department of Commerce under Grant BS123456".

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

case study, and its working environment. Then, we will demonstrate through this case study the process carried out at each step of our approach.

II. Theoretical Background

A. Domain Driven Design

Domain driven design is one of the pillars of our approach to address the complexity issue of legacy applications. The complexity may have two facets: the technological one and the application's domain of activity one. Reducing the complexity of the application's domain of activity, make possible to build cleaner and more sustainable applications. Moreover, this concept will enable the company to focus more on its strategy.

Domain driven design is about discussing, understanding and discovering business value, while aiming to centralize knowledge. It is centered on the application's business and the source code that implements it.

This approach has given birth to mix the different design and development practices, and to focus on how to cooperate the design part and the development part to create a better solution.

The principles of this approach are as follows [16]:

1) Focus on the core domain and domain logic.
2) Base complex designs on models of the domain.
3) Constantly collaborate with domain experts, in order to improve the application model and resolve any emerging domain-related issues.

B. Smart use cases

According to [17], this concept comes to answer many issues in a software development project, which is characterized by functional and technical highly complex. In order to mitigate this complexity, the project would be expressed using a single of unit work, which in turn is expressed in the smart use case.

By and large, the functional requirement of the project is modeled using traditional use cases, from defining the users' objectives to execute the steps. However, the traditional use case may vary considerably in its size and its complexity in terms of identification and realization of services. This makes the traditional use cases hard to specify, so in a result, it will be hard to implement and hard to test too. As an example, a use case may be required a hundred pages, a hundred scenario and many services to be described. So for designing the smart use cases, the use cases are split up into different levels of granularity in terms of size and complexity, such as:

1) Cloud: regroups clusters of the different business process that belong together.
2) Kite: Individual business processes are generally placed.
3) Sea: In this level, each single use case describes a single elementary business process and achieves a single goal.
4) Fish: This level is used to model autonomous functionality supporting the sea level.
5) Claim: the processes often need to be deeply modeled appeared as sub steps in its up level.

III. Agile Approach to Modernization and Migration of Legacy Applications to Native Cloud Applications

In this section, we propose the cycle of an agile modernization (see figure 1) approach, which focus to transform the legacy applications to native cloud application [18]. This approach is based heavily on the ADM approach by following all good modeling practices, in particular, domain driven design and smart use cases. This cycle consists of three major phases, as follows:

4) Recover PIM model
5) Transformation PIM source to PIM cloud
6) Transformation PIM cloud to PSM cloud

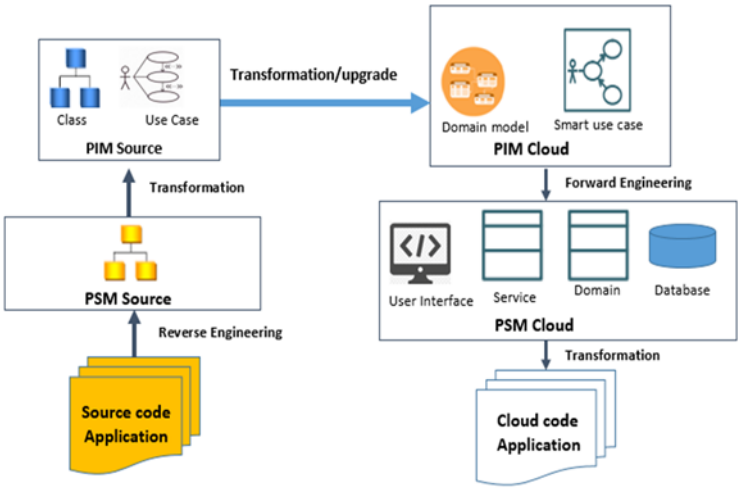


Fig. 1. Agile Approach to modernization and migration to the cloud

A. Recover PIM

The first phase of this approach uses reverse engineering to discover abstract models (PSMs), representing the existing platform of the application, and then obtaining its PIM model from these models. The objective of this first phase is, to express the application PIM model in two UML diagrams, in particular the use case diagram and the class diagram. The goal is to describe the business processes carried out by the application, and as well as its behavior and structure independently of any specific technology and platform. This phase consists of analyzing the system to identify its components and its relations with each other, and to design new representations of the system in another form or at a higher level of abstraction. The figure below explains the process to be followed to carry out this first phase:

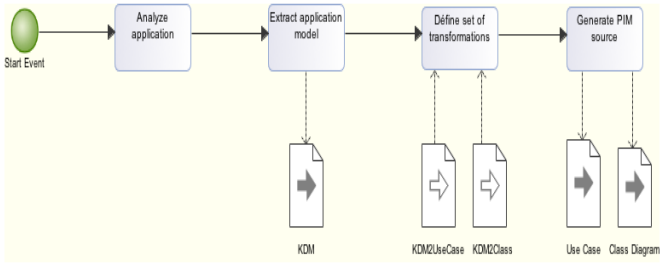


Fig. 2. Process Recover PIM source

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

1) Source code to KDM

The first sub-stage is about to reverse engineering a code, which aims to recover the model code of the application, and to create PSM conformed to KDM metamodel. This sub-stage is supported by means of code-to-model transformation. The PSM resulted is a model tailored to specify a system in terms of a specific platform, i.e., JAVAEE, expressed in Unified Modeling Language (UML). The subset of UML diagrams that are useful for PSM, includes the class diagram and the state diagram.

2) KDM to PIM source

In the second sub-stage is about to transform the KDM into PIM-UML models by means of a model to model transformation, that takes as an input a model conforming to the KDM metamodel and produces as outputs models conforming to the UML metamodel.

By doing so, we have proposed two kinds of model to model transformations, in a way the source metamodel corresponds to the KDM metamodel and the both targets metamodels correspond respectively to the UML metamodel of use cases and UML metamodel of a class diagram. Those two model-to-model transformations specified, are represented as follows:

--The first model to model transformation, called KDM2UseCases transformation, produces use cases diagram as the first target model from the KDM source model.

--The second one, called KDM2ClassDiagram, produces a class diagram as the second target model from the KDM source model.

B. Transformation PIM source to PIM cloud

The second phase is about to transform the source PIM into a cloud PIM, applying a set of activities to be transformed into a native cloud application. This cloud PIM will be designed in such a way that each basic business process operates independently. This cloud PIM should conform to the metamodel [19] [20]. In addition, this cloud PIM can be implemented on any cloud platform.

For this purpose, this phase includes two inputs generated from the previous phase: class diagram and use case diagram

1) Extract domain models

Extract different domain models of the application by using the UML class diagram. Each package will be represented as a set of classes, describing all the entities in a given domain. The objective of this sub-stage is to mitigate the complexity of the application by structuring it into distinct domains with a well-defined bounded context, where a domain model is defined and applicable.

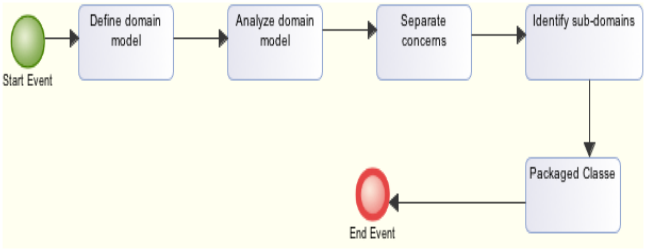


Fig. 3: Identification the sub-domains models

This sub-stage is generally carried out in four steps, namely:

--Define the domain model: The purpose of this step is to define the domain model of the application whose importance is to understand the business domain. The application domain model is represented by the class diagram classes, and its associations represent the relationships between them.

--Analyze Domain: This step involves understanding the domain of application across the relevant classes and its relationships to each other. However, each class in this diagram represents a conceptual class of the domain.

--Separate problems: By analyzing the application domain model in the previous step, and following the best practices of domain driven design, the entire application domain is composed of sub-domains. Each of which responds to a well-defined concern for a well-defined team.

--Identify the sub-models of the domain: In this step, despite focusing on developing a single model of the domain, which is generally extremely complex and overly broad, we separate classes by applying the bounded context concept to define sub-domains with a clear and visible context definition.

--Packaged classes: Finally, in this step, we will regroup all classes and their associations, which represent a specific sub-domain.

C. Identify smart use cases

A use case describes the application's functionality and the processes to be performed. Each use case presents a feature and describes a set of actions to be done. The smart use case will be identified as sub-functions of its use case mother. Each one represents an elementary business process. This sub-phase takes place in 5 steps guaranteeing the functional decomposition of the application into autonomous and independent functionalities.

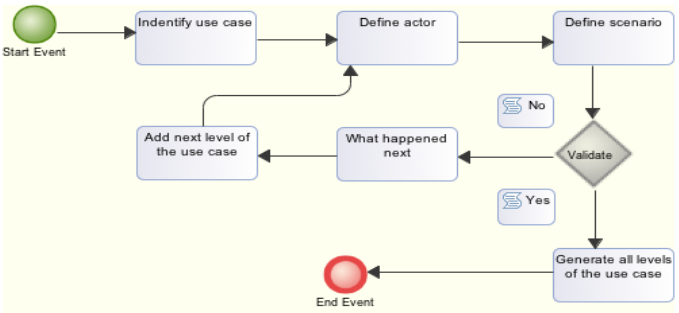


Fig. 4: Generation process of the smart use cases

Hereinafter the various stages of sub-phase:

1) Identify use case: In this step, we will break down the use case diagram from the previous phase to separate use cases. Generally, a use case defines a way of using the system and describes its functional requirements. Each use case will be modeled as a single business process.
2) Define actors: An actor is anyone or something with behavior, to achieve a specific goal. In this step, the actor should be defined to realize the use case.
3) Define Scenario: A scenario describes a sequence of steps

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

that are executed in order to achieve the purpose that the use case is supposed to deliver. In this step, the scenario will define a single elementary business process of use case identified in step 1.

4) Validate: After defining the use case scenario, this step consists of validating the scenario conformity with the expressed need. This will make possible to judge if there is a need to add another level of granularity of use cases, at the Fish level. This step is done by continually asking the following question: what is the following step that we need to take in performing this use case? Sometimes, it is useful to record these steps as individual use cases in the diagram, and sometimes this is not the case.
5) Generate Use Case Levels: Finally, in this step, use case levels added to the parent use case will be generated, constituting a single, basic business process of this use case.

It should be noted that the steps of this sub-phase are iterative in order to complete all the cases of use of the initial diagram. In addition, when complex scenarios arise, sequence diagrams can be used to further detail the use cases including the user's goal.

Finally, to complete the transformation phase, each elementary business process should be associated with its sub-domain to which it belongs. To this end, in every basic business process, smart use cases will be mapped. In turn, the sub-model of the domain represents this business process will be configured, in order to keep the link of belonging to its smart use cases to its own domain model.

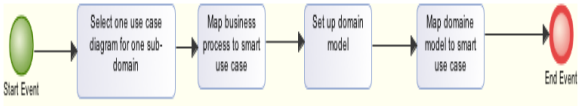


Fig. 5: Associate smart use case to its sub-domain

D. Transformation PIM cloud to PSM cloud

The third phase uses forward engineering techniques, which consist of a set of activities that produce a native cloud application by applying a series of well-defined transformations.

In this context, this phase will present how to transform the cloud PIM model into a set of services deployed in a cloud platform.

The cloud PIM model consists of a set of smart use cases, to be transformed into a set of services and deployed in a cloud platform.

In this step, we present how to transform this cloud PIM model into a cloud platform. Every smart use case in the cloud PIM will be implemented as follows:

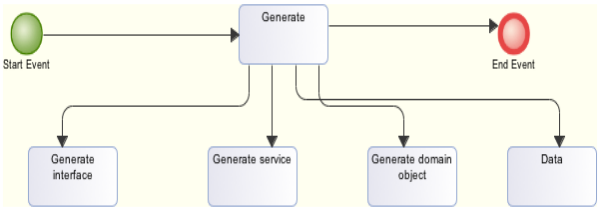


Fig. 6: Implementation smart use cases

1) A smart use case is implemented independently, allowing the application to be scalable and deployed quickly.
2) Each smart case manages the navigation separately, which guarantees the support of a wide variety of user interfaces.
3) Each of the smart use case performs individual tasks that meet a very specific business scope.
4) Each of the smart use case has its own domain that ensures the relationship between smart use case and its domain (package).
5) Finally, each of the smart use case has its own database which allows to have different types of storage in the cloud.

IV. Case Study

A. Reference application

Throughout this article, the application of a mail management shown in Figure 7 is used as running example of transforming a legacy application into a cloud native application. It is a complex, multi-stakeholder system that lends itself to domain driven design. This case study will be the object of the modernization and the migration to the cloud of a following system respecting the proposed approach. We focus the study on the transformation of the PIM, obtained through the phase of the inverse engineering, to a cloud PIM according to our approach.

This case study application consists of three layers: the presentation layer, the service layer, and the data access layer:

1) The presentation layer

It is used in human-machine interactions. It is represented in the mail management system the IHM part developed by the RIA solution.

2) The service layer

It has two roles: orchestration services, and business services.

3) The persistence layer

This layer is responsible for the persistence of data through a database. The data of the application were centralized in a single database of HSQL (Hyperthreaded Structured Query LanguageDatabase)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

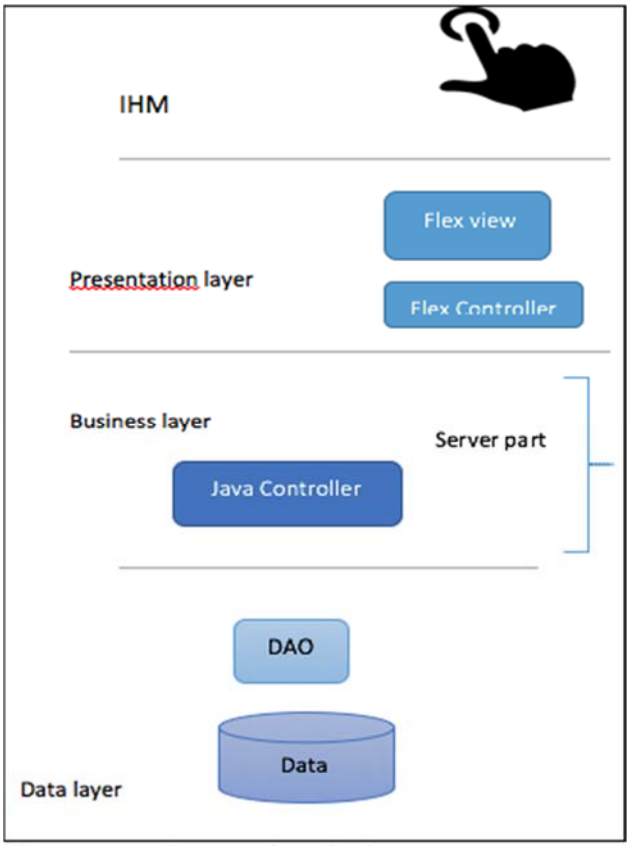


Fig. 7: Architecture of application to be moved to cloud

B. *Validation of the approach*

1) *Recover PIM*

This phase follows the techniques of reverse engineering to discover the abstract KDM model representing the existing platform of the application used and then developing its PIM model from the KDM model. Concretely, we have discovered the KDM model using the Modisco tool that implements the Java metamodel. This tool makes it possible to produce the application's KDM model from its Java code, by deducing its global structure (for example, package, classes, dependencies) using the code-to-model transformation, thus model-to-model transformation the PIM model from the KDM model.

As described previously in the approach, this phase comprises two sub-phases:

--Source code to KDM: it is realized via the Modisco tool, which has for the purpose to extract models through the code of the application. In this case, it is a KDM class diagram of the application written in Java. KDM provides a coarse vision of application components, such as classes, functions, or data, as well as the relationships that unite them.

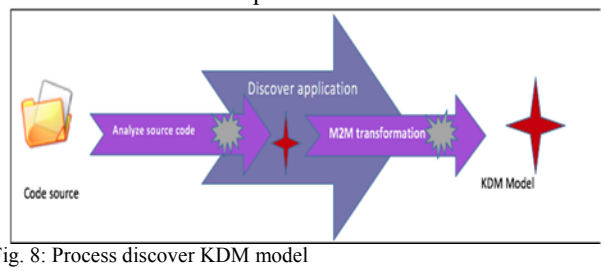


Fig. 8: Process discover KDM model

--KDM to PIM: is the transformation of the KDM model into UML models at the PIM level. In this step, the application's KDM model, retrieved from the source code transformation to KDM, is used as the starting point to retrieve PIM templates using KDM-to-UML transformations. This transformation has been implemented as a model-to-model ATL transformation that takes as input a model conforming to the KDM metamodel and outputs a model conforming to the UML metamodel. In this sense, two types of model transformations have been proposed:

- **KDM2UseCases transformations**: The KDM2UseCases module that corresponds to the transformation specifies how to produce use case diagrams (target model) from the KDM (source model) model.

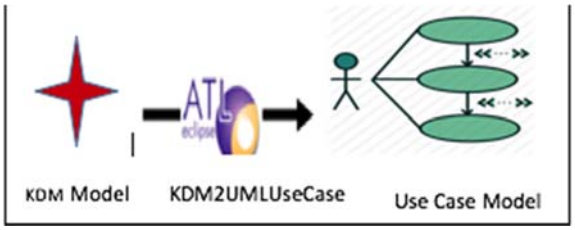


Fig. 9: Transformation KDM to use case model

As example of the ATL module transformation of KDM model in use case model, we choose the transformation of KDM package to UML package, we obtain:

```
RuleKDMPackage2UMLPackage {

   from p: MM!Package

   to UMLPackage:MM1!Package (

         name <- p.name,
packagedElement <-p.get_MethodUnit()

) }
```

This transformation explains how to transform KDM package to UML package

- **KDM2ClassDiagram transformation**: This transformation module specifies how to produce class diagrams (target model) from the KDM model (source model). The source and target models must conform to their respective metamodel.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig. 10: Transformation KDM to class diagram model

As example of the ATL module transformation of KDM model in class diagram model, we choose the transformation OperationKDM2OperationUML, to transform each KDM method to UML method, by remove all platform specificities in the source model. In this case, the rule has for the purpose of removing all the Setters and Getters from the application. So we obtain:

```
Rule OperationKDM2OperationUML {
from OperationKDM:uml!Operation (
not (s.Getters()='get') and not(s.Setters()='set'))
to  operationPIM:uml!Operation (
        name<- OperationPSM.name,
        visibility<- OperationPSM.visibility,
      ownedParameter<- OperationPSM.ownedParameter
   ) }
```

The two following figures illustrate the results of the two previous transformations:

– The Use Case Diagram: This diagram describes the functionality of the application and the business processes to be performed. Each use case presents a feature and describes a set of actions to be performed.

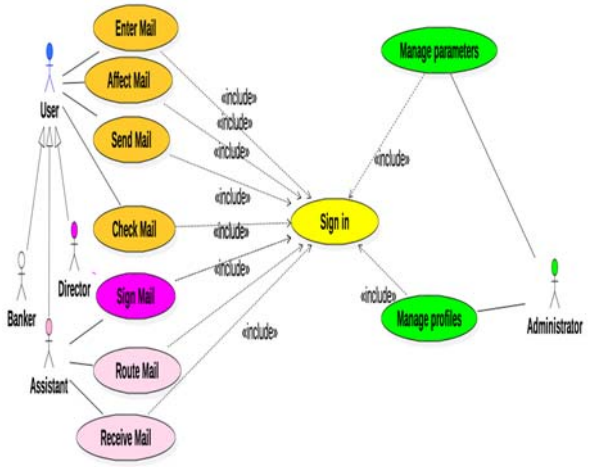


Fig. 11: Target model resulting from the KDM2UseCase transformation

A. The class diagram: This diagram is for the purpose of helping domain experts analyze the application domain model, to use domain vocabularies, to understand the conceptual model class of the application.

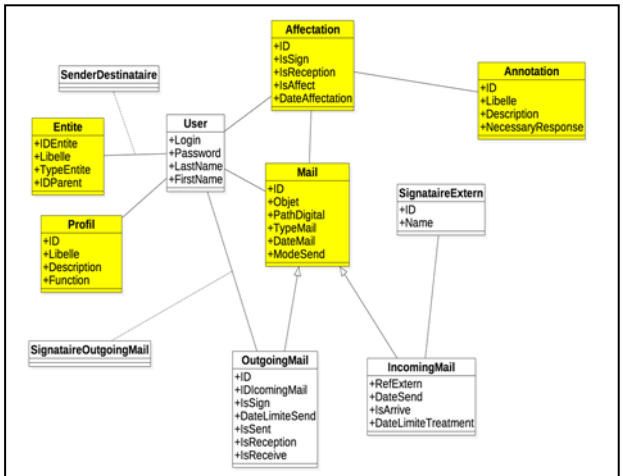


Fig. 12: Target model resulting from the KDM2ClassDiagram transformation

The domain model, from this class diagram, includes the classes, are:

- The Courier class
- The Profile class
- The Entity class
- The Assignment class
- The Annotation class

2) Transformation PIM to PIM cloud

-- Extract domain models:

Extract the domain model using the class diagram. Due to the need to separate application concerns by a domain, each package will be represented as a set of classes belonging to the same domain, describing all the entities of a given domain.

In this context, we note that the domain model of the application consists of these following sub-domain models:

1. Mail Management Domain
2. Mail Processing Domain
3. Parameter management Domain
4. Profile Management Domain

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

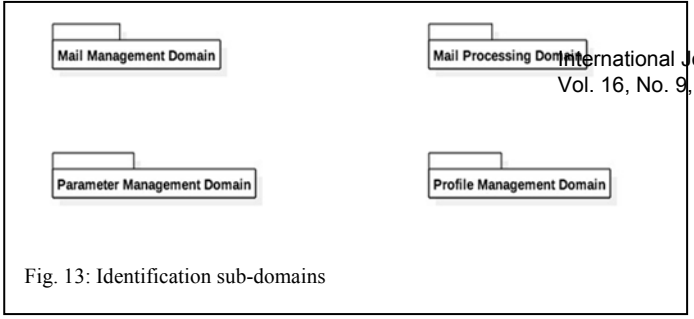


Fig. 13: Identification sub-domains

-- Identify smart use cases

Using the use case diagram of the application, which describes its functional requirements and the processes to be performed, smart use case will be identified as sub-functions of the parent use case, and each will represent an elementary operational process. As described above, smart use cases have been identified according to the granularity levels of the use case.

In this case study, the scenario that describes how "enter a mail" might look like this:

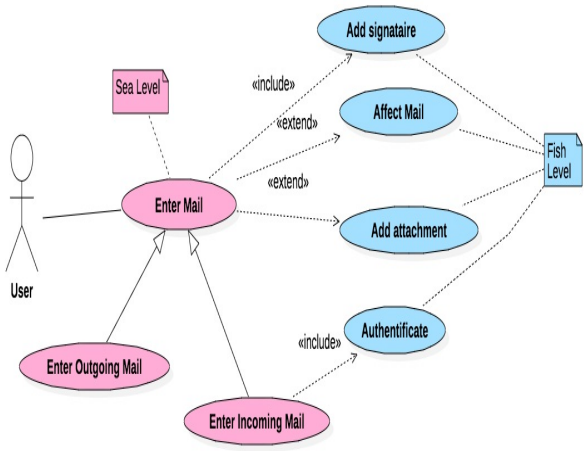


Fig. 14: Smart use case "Enter Mail"

The smart use cases are defined at Sea level and at the Fish level, in this case:

- Sea level: Enter Mail
  - Enter outgoing mail
  - Enter incoming mail
- Fish level: Add signatory
  - Assign mail
  - Sign in

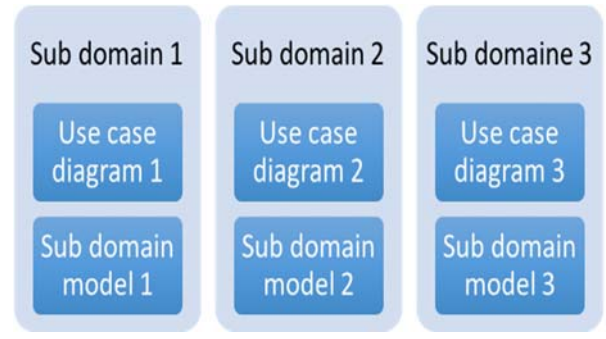


Fig. 15: Association Smart use case to sub-domain

- Add attachments

This use case of "Enter Mail" belongs to the domain model "Mail Management".

3) Transformation PIM cloud to PSM cloud

In this phase, we have redesigned the application by applying the previously defined activities, which is how to generate smart use cases.

What follows the final architecture of transformation from the legacy application to the native cloud application, applying the set of transformations to cases of smart uses of each sub-domain of the application.

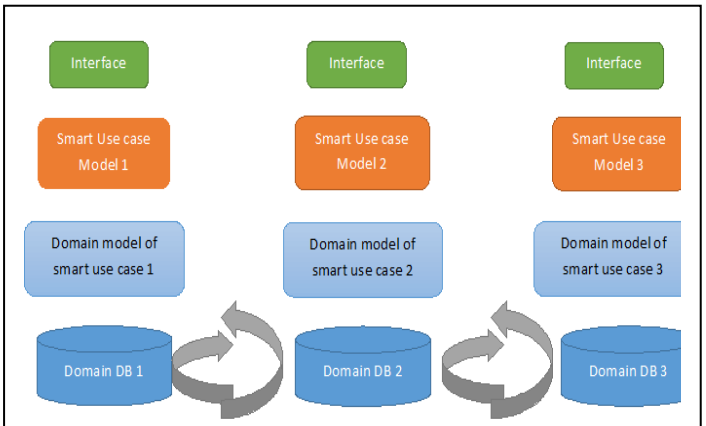


Fig. 16: New architecture of migrated application

This architecture allows us to have the following results:

1. The application has become modular: each team has its own full lifecycle of its core business.
2. Segmentation of complexity: the domain driven design concept has lead to segmentation in sub-domains to separate the preoccupations of each stakeholder. This process independently deals with its own problems in its own sub-domain.
3. Scaling: application modules are now scaled independently, to meet the increasing workload, the number of resources assigned to a client or application increases.
4. Quick and easy deployment: Instead of deploying the application entirely, this architecture allows to deploy services independently and fast way, while each service is autonomous and doesn't depend on anything

V. CONCLUSION

The agile approach of modernizing legacy applications to the native cloud applications that we have presented in detail throughout this paper, presents a new path to take advantage of all the benefits of cloud. This approach is extensible by favoring the addition of new business functions in a modular way. In addition, the agile approach follows an iterative and incremental process of modernization to support this modernization. The purpose of this process is to make the application more agile to ensure business continuity requirements and the longevity of IT. Through this process,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the application can be scaled modularly, and application components, which means smart use cases, have been implemented, deployed, configured, distributed and maintained in an independent way, taking profit from all the benefits of the cloud.

We propose the validation of this approach by a case study by choosing as a native application a mail management application. It is a complex system involving several actors. The main interest of this case study is to show how the PIM model of the application will be transformed to a cloud PIM, and meets all the definitions provided for the native cloud application.

As a future work, we want to use this agile approach and make it a standard that applies to any project of modernization of the legacy application to native cloud application.

REFERENCES

[1] Q. Zhang, L. Cheng, R. Boutaba. Cloud computing: Stateof-the-art and research challenges. Journal of Internet Services and Applications, vol. 1 no. 1, pp. 7–18, 2010.

[2] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic. "Cloud computing and emerging IT platforms: vision, hype, and reality for delivering Computing as the 5th Utility. Future Generation Computer Systems. 2009.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I.Stoica, M. Zaharia. Above the clouds: a Berkeley view of cloud computing. technical report. 2009

[4] S. Jha, A. Merzky, G. Fox. Using clouds to provide grids with higher levels of abstraction and explicit support for usage modes. Concurrency and Computation: Practice and Experience. 2009

[5] National Institute of Standards and Technology. The NIST definition of cloud computing. 2009.

[6] Matt Stine, Migrating to Cloud-Native Application Architectures Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. 2015

[7] Ardagna, D., di Nitto, E., Mohagheghi, P., Mosser, S., Ballagny, C., D'Andria, F., Casale, G., Matthews, P., Nechifor, C.S., Petcu, D., Gericke, A., Sheridan, C. Modaclouds: A model-driven approach for the design and execution of applications on multiple clouds. 4th International Workshop on Modelling in Software Engineering (MISE). 50-56, 2012.

[8] Bergmayr, A., Bruneliere, H., Canovas Izquierdo, J., Gorronogoitia, J., Kousiouris, G., Kyriazis, D., Langer, P., Menychtas, A., Orue-Echevarria, L., Pezuela, C., Wim- mer, M. Migrating legacy software to the cloud with artist. 17th European Con- ference on Software Maintenance and Reengineering (CSMR), 465-468, 2013.

[9] Fehling, C., Leymann, F., Ruehl, S., Rudek, M., Verclas, S. Service migration patterns: decision support and best practices for the migration of existing service- based applications to cloud environments. 6th IEEE International Conference on Service-Oriented Computing and Applications (SOCA), 9-16, 2013.

[10] Jamshidi, P., Pahl, C., Chinenyeze, S., Liu, X. Cloud migration patterns: A multi- cloud architectural perspective. 10th International Workshop on Engineering Service-Oriented Applications (WESOA). 2014.

[11] Mendonca, N. Architectural options for cloud migration. Computer 47(8). 62-66. 2014.

[12] Fehling, C., Konrad, R., Leymann, F., Mietzner, R., Pauly, M., Schumm, D.,Flexible Process-based Applications in Hybrid Clouds. IEEE International Conference on Cloud Computing (CLOUD). 2011

[13] Brandic, I., Anstett, T., Schumm, D., Leymann, F., Dustdar, S., Konrad, R. Compliant Cloud Computing (C3): Architecture and Language Support for User-driven Compliance Management in Clouds. the 3rd International Conference on Cloud Computing (Cloud). 2010

[14] K. Sabiri, F. Benabbou, M.hain, H. Moutachouik, K. Akodadi, A survey of cloud migration methods : A comparaison and proposition, doi: 10.14569/IJACSA.2016.070579 (ijacsa), Volume 7 Issue 5, 2016.

[15] Eric Evans. Domain Driven Design - Tackling Complexity in the Heart of Software. 2003

[16] Vaughn Vernon, Implementing Domain Driven Design. 2013.

[17] Khadija Sabiri, Faouzia Benabbou, Adil Khammal. Model Driven Modernization and Cloud Migration Framework with Smart Use Case. The Mediterranean Symposium on Smart City Applications, 2017.

[18] Mohammad Hamdaqa, Tassos Livogiannis and Ladan Tahvildari, A reference model for developing cloud applications, In Proceedings of the 1st International Conference on Cloud Computing and Services Science, pages 98-103, DOI: 10.5220/0003393800980103, CLOSER 2011

[19] Khadija Sabiri, Faouzia Benabbou, Hicham Moutachouik, Mustapha Hain: Towards a cloud migration framework, Third World Conference on Complex System (WCCS), 2015

SABIRI Khadija was born on the 7th of March, 1989 in Casablanca, Morocco. She recieved his Ph.D in Computer Science from Science Faculty of Ben M'sik, University Hassan 2. his research aims to develop a process of modernization and migration from a legacy application to native cloud application. His current research area focuses on cloud computing, smart parking as a part of Internet Of Thing, machine learning.
Email: Khadija.sabiry@gmail.com

Faouzia Benabbou is an professor of Computer Science and member of Computer Science and Information Processing laboratory. She is Head of the team "Cloud Computing, Network and Systems Engineering (CCNSE)". She received his Ph.D. in Computer Science from the Faculty of Sciences, University Mohamed V, Morocco, 1997. His research areas include cloud Computing, data mining, machine learning, and Natural Language Processing. She has published several scientific articles and book chapters in these areas.
Email: Faouzia.benabbou@univh2c.ma

Mohamed amine Hanine, PhD student in IR2M laboratory at the Faculty of Science and Technology of Settat, university HASSAN 1st Morocco. He is consultant, working on reengineering (Cloud migration) and software quality.
Email: m.a.hanine@gmail.com

Khalid Akodadi, received the Ph.D. degrees in applied mathematics and

computer science applied to artificial intelligence from the Faculty of Science Ben M'Sick, Hessan II University, Casablanca, Morocco, in 2009, he is currently preparing a thesis titled "Semantic middleware: management of audiovisual broadcasting in Big Data with dematerialized digital television channel" at Laboratory LAROSERI in the Faculty of Science, Eljadida, Morocco. Her current research interests include neural networks, Artificial Intelligence, Machine Learning, Clustering, Regression and Classification, Deep Learning, Business Intelligence, Text Analytics, Natural language processing, Business Process Models, Architecture of Cloud, Big Data Analytics.
: khalidakodadi@gmail.com

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A novel Item Recommender for mobile plans

Neetu Singh
dept. of Computer Science
Mody University of Science and Technology,
Sikar, India
nits.tanwar29@gmail.com

V.K Jain
School of Engineering and Technology
Mody University of Science and Technology,
Sikar, India
dean.cet@modyuniversity.ac.in

***Abstract*—With constant enlargement of the scope and coverage of mobile market, the traditional algorithms for short message service (sms) just help to tell all schemes related to particular chosen network, without considering the needs of particular individuals. Therefore, various recommender systems commissioning different data representations along with recommendation methods are presently used to cope with these challenges. A original scheme for item-based recommendation is proposed in this paper that just focuses on user interest as top most priority, while filtering superfluous messages. This recommender in turn enhances user experience regarding their selected network. Progressive experimental results are provided to showcase the usefulness of our method.**

Keywords-: Recommender Systems; Cellular networks; Similarities; data plans

I. INTRODUCTION

The exponential evolution of the world-wide-web along with the hype of e-commerce has led to the expansion of recommender system .It is a personalized information provider to recognise item sets that will be of interest to a particular user. Recommender systems are the base for the future of the smart webs. The systems produce smooth user experience by making information retrieval easier and divert users from queries typing phase towards hit it off suggested links. No one is untouched by real-life recommender systems .They are doing amazing work, when browsing for music, movies, news or books. These engines are mandatory for websites like Amazon, Myntra or Netflix. In fact the core of recommender system is its algorithm, it has been characterize into four methods, namely graph model recommendation, collaborative filtering recommendation, hybrid recommendation and content-based recommendation [1]. On the basis of different approaches used for development of recommender systems such as demographic, content, or historical information [2,3,4] , User based collaborative filtering came out as the most popular and promising one for building recommender systems to date .Yet it is most successful but unluckily, the linear growth of its computational complexity with the count of customers can grow to be several millions as witnessed in commercial applications. To overcome this scalability issues, the item-based recommendation techniques have been developed .It generates user-item matrix to recognize relations among the diverse items, and use them to produce recommendations [5,6].Forming the clusters of user is one of the solution for reducing the complexity of this nearest-neighbour computations or use the cluster centroids to originate the recommendations [7,8].At Amazon.com, the item recommendation algorithms is used to create the personalize online store for every customer. This store fundamentally changes according to customer interests, showing a software engineer, the title of programming and baby accessories to a mother[9].

In this paper we present an item recommender for mobile plans that precisely hit customer interest and hence recommends tariff, roaming, data and top up plans of his interest instead of telling all basic plans to him.

The rest of this paper is organized as follows: section 2 will outlines the proposed Architecture. The experimental results have been discussed in section 3, finally, the conclusions and future work in section 4.

II. MOBILE PLANS ITEM RECOMMENDER ARCHITECTURE

For the mobile item recommender, the architecture is as shown in Figure 1. It has five phases. Let's discuss the architecture in detail:

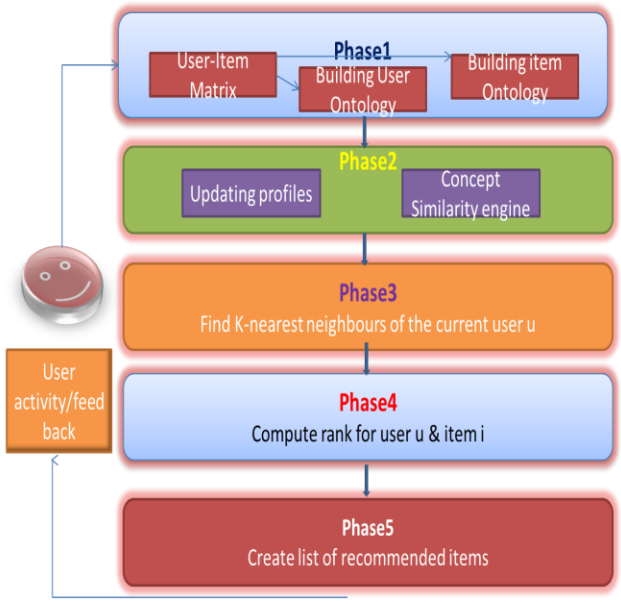


Figure1: Architecture of mobile plans item recommender

A. *Phase 1:*

Phase1 is all about creating user-item matrix and considering item & user ontology. Data creation is very important in RS. Some of the techniques consider only basic

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

information like rankings or one can say ratings while others act on extra knowledge like social as well personal constraints, and real time activities in case of distributed systems.

Basically for recommendations, there should be something common between users or the items. That means focus is on item, user and the tractions between them. In our item recommender, we have preferences; these are in form of user ID, item ID and the user preference for the item.

As our recommender is based on mobile plans, we have chosen 'Airtel' as the network and its available plans like talk time, data, special cutters, SMS as an item for our item recommender. In user ontology, we have users using mobiles, divided into two categories men and women. As supported by general fact, men and women are completely different. Former like gaming and watching live cricket match while latter focuses on online shopping, calling, which in turn clearly affect the preferred mobile plans. Other information regarding users are locality, state, annual income, occupation, age and their general usage time for data using general features like web, call, SMS etc. The item ontology has item, item id, and rating of particular item as given by user for utility as pack. Table 1 gives a short description of item, item id and its description [14]. This all is then passed to the second phase.

TABLE 1: Airtel mobile plans

Cate-gory	Plan Details of Airtel					Item ID
Data Packs	2G	3G	4G	Other benefits	Days	
	90MB				3	10
		250 MB			28	17
		1.25GB		Wynk plus internet	28	25
Mixed	Data	STD	Local	Landline	SMS	
	110 MB 2G	2.5p/s	2p/s	2.5p/s		68
	7GB 3G	Unlimited	Un-limited	unlimited		75
						1Q
FC	Talktime	Incoming	Out going	Minutes		
				85		92
	1479	20p/s	30p/s			97
SMS	Local	National	Intern-ational	Days		
	150	175		28		100
	300	300		28		110
Tarriff	Minute's local	Minute's local	Local call	STD call	Talk time	
					1050	159
	90					170
			40p/s	40p/s		200

B. *Phase 2:*

After completion of phase1, we got a clear idea about what parameters are suitable for getting similarity measures between items and users. So we need only that information that is meaningful and could be passed to recommender. This is basically updating of profiles.

There are various ways to define similarity measure between two vectors x = (x1; : : : ; xn) and y = (y1; : : : ; yn). Mahout Library facilitates us with similarity models namely: Pearson correlation similarity, Euclidean distance similarity, Spearman correlation similarity, Cosine measure similarity and Tanimoto coefficient similarity. Last one is set of operations generally represented by binary value and is effective only when user expressed some preference for an item. Pearson similarity and cosine similarities are quite related to each other. Major difference between two is that cosine similarity works on centred data. Euclidean distance between two vectors became Euclidean Similarity when we get negative proportion to this distance. Spearman similarity is same as Pearson correlation similarity except that value xi and yj are replaced with their relative ranks. The detailed explanations can be found in the javadoc API of Mahout [10].

C. *Phase 3:*

In this phase we are going to calculate the 'n' neighbour of the current user 'u' and pass size of neighbourhood as parameter in Item based Recommender. One of the KNN algorithms is used for the purpose. Size of the neighbourhood is important for the accuracy of the algorithm. The utmost similar users are taken (the number is limited by the neighbourhood size).

D. *Phase 4:*

Finally recommender generates some ranks based on preferences and similarities. The algorithm of similarity for Item recommender is as follow [9]:

For every item I1 in item ontology
For every user U in user ontology who bought I1
For each item I2 bought by
user U
Record that a user bought both I1
and I2
For each item I2
Calculate the similarity between I1 and I2

E. *Phase 5:*

A list of most suitable items for recommendation has been generated and passed to user. Good recommenders suggests best suited items to user but how good is our recommender, it depends upon performance evaluation. This has been discussed in next section of result and discussion.

III. Results and Discussions

For checking how good is the recommender? Various experiments had been performed on the considered dataset to check: if item recommender is able to suggest the current plans to the users or not. The different datasets used in our experiments are as follows:

I. 50,000 users generating 1.5 lakhs of data

II. 70,000 users generating 2 Lakhs of data

III. 100,000 users generating 3 Lakhs of data

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

According to machine learning, it is very important to train your model and then test it on new datasets. For our experiments, dataset had been divided into 80% and 20%.Former for training and later for testing. Some of parameters used for our research are region, annual income, designation, locality, and the most frequently used internet hours in a day. But prime parameter is Item ratings given by users for their used packs. All this data is passed to the Item Based recommender, where it processes data and generates the similarity metrics and finally creates the list of most suitable items for the particular users according to their need. i.e. there is quite bright chance that user appreciates the plan and starts using it:

A. Evalvation

As we are aware, that recommender quality is directly proportional to accuracy. Our focus is to determine how accurate is the item recommender? There are two perspective methods for accuracy measurement namely statistical and decision. There have been many evaluation metrics for recommendation algorithms, such as recall[11], mean absolute error, mean square error [12],NMAE[13].

Let's consider two of the statistical method and some of the decision methods. Based on these methods: Graph A, B, C, D are the representation the of the evaluation of our item based recommender on the considered data sets. Results show how well it suggests, the best suited mobile plans to the particular customer while stopping all unwanted messages.

- Mean absolute error (MAE): It is the measure of recommendation deviation from user specific value. Its mathematical formulae is as follows:-

$$\mathrm{MAE} = \frac{1}{N}\sum_{u,i}|p_{u,i} - r_{u,i}| \quad (1)$$

Where,
$p_{u,i}$ = Predicted rating for user u on item i,
$r_{u,i}$ = Actual rating
N = Total number of ratings on the item set.

MAE is inverse proportional to recommender performance

- The Root Mean Square (RMSE): It is related to absolute error. "RMS gives the difference between actual and estimated value in testing set example". In current experiment, the ratings range from 1 to 5. RMSE formulae is:-

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{u,i}(p_{u,i} - r_{u,i})^2} \quad (2)$$

It is also inversely proportional to recommender performance.

Decision support metrics:

- Precision:" Precision is the fractional measure of suggested items which are of real interest to the user. Precision support accuracy metric is the metric that helps in selecting high quality items from available set of items."[14].More the precision, better the recommendations. It ranges from 0 to 1.

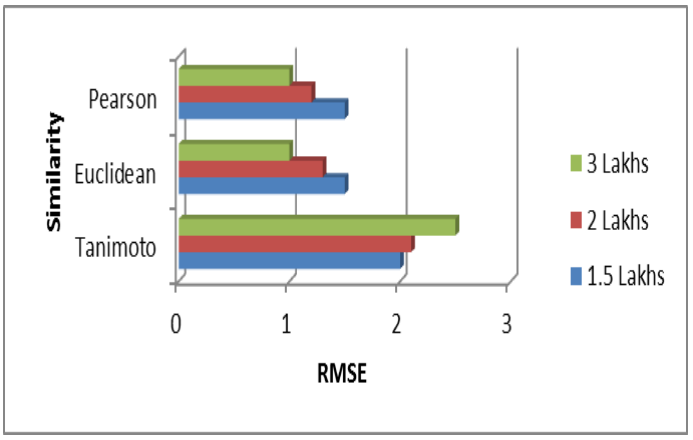
$$\mathrm{Precision} = \frac{\text{Correctly recommended items}}{\text{Total recommended items}} \quad (3)$$

- **Recall**: "fractional measure of relevant items that are also part of set of suggested items is known as Recall.[14]".Lesser the recall value , better are the recommendations.

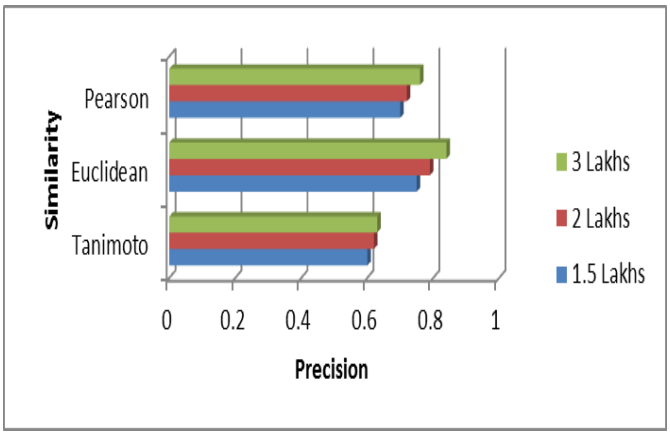
$$\mathrm{Recall} = \frac{\text{Correctly recommended items}}{\text{Total useful recommended items}} \quad (4)$$

- **Fall out**: It is calculated with the combination of both i.e precision (P) and recall(R). Formulae is

$$F\text{-measure} = \frac{2PR}{P+R} \quad (5)$$



Graph A:Similarity Vs RMSE

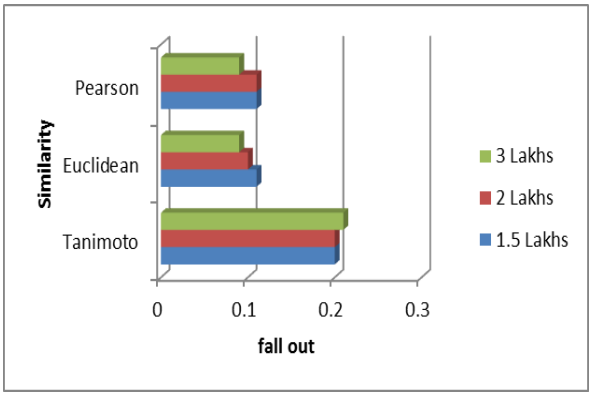


Graph B:Similarity Vs Precision

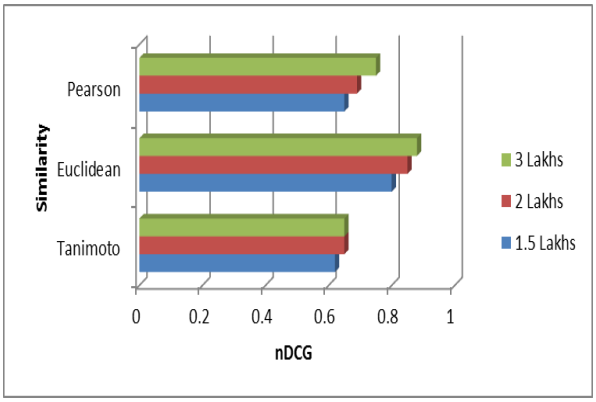
- **nDCG:** It is a fact that recommender suggests some items based on some relevance core, basically a non-negative number, called a gain in nDCG. In case, no

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

user feedback is available, gain is set to zero. By adding all gain, we get cumulative gain. It is used to put most exact items on top. So, before calculating these courses every individual score is divided by increasing number (generally log of position of item) known as discounting (DCG). But, for using DCG's for comparison between users, we need to stabilize them. It's value ranges between 0 to 1.It is directly proportional to good recommendations.



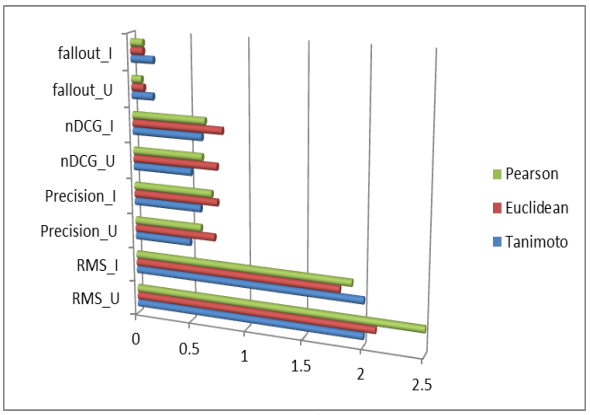
Graph B:Similarity Vs Fall out



Graph B:Similarity Vs nDCG

From all these experiments, which are implemented recommender algorithms through the Mahout java library (an open-source implementations of machine-learning algorithms. It is powering several portals e.g. Yahoo! Mail and AOL).We have observed that latency of Item Based model is very much dependent on the selection of similarity measure. The fastest recommendations are output of Euclidean and Pearson similarities. Tanimoto is the slowest and does not scale well with the density growth.

From the Graph E, when the data is same that is users are same as well as items are same[in this case we have taken dataset 1i.e50,000 users generating 1.5 Lakhs data], we have found that the performance of item recommender is better than user based recommender[14].Hence, we concluded the item recommender is better than User based recommender, at least in suggesting right mobile plan to the user.



Graph E: Comparison Graph of User based & Item based Recommender

IV Conclusion and Future Work

In this paper we propose a novel item based recommendation method which is capable of distributing right mobile plan to the right user. It is beneficial to both, the user of network and for the network providing company. We successfully implemented it and results show that among three methods of similarity calculation, Ecludien and Pearson perform better than Tanimoto in sorting accuracy and hitting rate.

In the future we plan to evaluate how the performance of recommender systems is altered when they are implemented in a distributed environment and how the recommender efficiency can be increased.

References

[1] Jianguo Liu, Tao Zhou, and Binhong Wang, Research process of personalized recommendation system, Progress in Natural Science,2009, 19(1): 115.

[2] Breese, J. S., Heckerman, D., Kadie, C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. UAI-1998.

[3] Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., Riedl, J. 1994. Grouplens: An Open Architecture for Collaborative Filtering of Netnews. SIGCSCW-1994.

[4] Schafer, J., Konstan, J., Riedl, J. Recommender systems in e-commerce. In Proceedings of ACM E-Commerce-1999.

[5] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143–177, 2004.

[6] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In Proc. of the WWW Conference, 2001

[7] Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa, and Jim Witshire. Discovery of aggregate usage profiles for web personalization. In Proceedings of the WebKDD Workshop, 2000.

[8] Lyle H. Ungar and Dean P. Foster. Clustering methods for collaborative filtering. In Workshop on Recommendation Systems at the 15th National Conference on Artificial Intelligence, 1998.

[9] Greg Linden, Brent Smith, Jeremy York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan./Feb. 2003,doi:10.1109/MIC.2003.1167344

[10] The mahout website." http://mahout.apache.org/.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[11] Deshpande, M., Karypis, G. Item-Based Top-N Recommendation Algorithms. ACM Trans. On Information Systems-2004.

[12] Marlin, B. Collaborative Filtering: A Machine Learning Perspective. Phd thesis. University of Toronto. 2004.

[13] Sarwar, B. M., Karypis, G., Konstan, J., Riedl, J. 2001. Item-Based Collaborative Filtering Recommendation Algorithms.WWW-2001.

[14] Ms. Neetu Singh, Dr. Puneet Kumar & Prof. Anil Kumar Dahiya, "RWYW: Recommend What You Want"-A Recommender for Mobile Plans, International Journal of Innovations & Advancement in Computer Science,Volume 7, February 2018

[15] J. Serrano-Guerrero, E. Herrera-Viedma, J.A. Olivas, A. Cerezo, F.P. Romero, A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0., Information Sciences 181 (9) (2011) 1503–1516.

[16] C. Porcel, J.M. Moreno, E. Herrera-Viedma, A multi-disciplinar recommender system to advice research resources in university digital libraries, Expert Systems with Applications 36 (10) (2009) 12520–12528.

[17] C. Porcel, E. Herrera-Viedma, Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries, Knowledge-Based Systems 23 (1) (2010) 32–39

[18] J. Bobadilla, F. Serradilla, A. Hernando, Collaborative filtering adapted to recommender systems of e-learning, Knowledge Based Systems 22 (2009) 261–265.

[19] Lathia N, Hailes S, Capra L, Amatriain X," Temporal diversity in recommender systems". In: Proceedings of the SIGIR 2010, July 19–23, Geneva, Switzerland, 2010.

[20] Li Y, Zhai CX, Chen Y," Exploiting rich user information for one-class collaborative filtering. Knowl Inf Syst 38", 2014, pp: 277–301.

[21] Bambini R, Cremonesi P, Turrin R,"A recommender system for an IPTV service provider: a real large-scale production environment". In: Ricci et al (eds) Recommender systems handbook. Springer, New York, NY, 2011.

[22] Singh and Reddy," A survey on platforms for big data analytics". Journal of Big Data, springer open journal, 2014 pp: 1-8.

[23] Blondel VD, Decuyper A, Krings G," A survey of results on mobile phone datasets analysis". EPJ Data Sci , 2015, pp :4-10.

[24] Miritello G, Rubén L, Cebrian M, Moro E," Limited communication capacity unveils strategies for human interaction". Sci Rep 3:1950, 2013.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Mathematical Knowledge Interpretation from Heterogeneous Data Sources– Industry Perspective

Savitha .C [1] Suresh Babu GOLLA[a,1]
[a] *Emerging Technologies and Solutions, EXILANT Technologies, a QuEST Global Company, Bangalore, India*
[1] *Suresh Babu Golla, Savitha Chinnappareddy, Emerging Technologies and Solutions, Exilant a QuEST Global Company, Bangalore, India; E-mail: sureshbabu.g@exilant.com, savitha.c@exilant.com .*

Abstract. In the current era, many industries and organisations are dealing with huge amount of data. Data plays vital role in many aspects and it is key to boost a company's revenue and fortune. Making data usable is a great challenge, as data exists in heterogeneous formats, syntax and semantics. Ontology is a prominent technology playing vital role in several application domains such as military, financial, insurance, manufacturing, and healthcare etc. and is widely used for knowledge interpretation and knowledge sharing between humans and machines. In industry perspective, this article describes ontology driven approach that overcomes semantic and syntactic heterogeneity in the data and interprets knowledge from the heterogeneous data sources such as Excel, csv, relational databases, key value stores and unstructured documents.

Keywords. Ontology, Data, Heterogeneity, Knowledge Base

1.Introduction

Ontology is a conceptualisation of a domain of interest [1]. It means that formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a domain of discourse. Ontologies enable the sharing of information between disparate systems within the same domain. There are numerous freely available ontologies from various industries. Ontologies are a way of standardising the vocabulary across a domain [2]. The standardisation allows for more flexibility and will enable more rapid development of applications and sharing of information.

Ontologies are used in many disciplines but are most commonly associated with Artificial Intelligence (AI), and possibly Natural Language Processing (NLP) applications. However, a less likely application of ontologies and an area of interest for us is with Data Integration and Knowledge Management systems. In linguistics, ontology refers to coexistence of possible meanings for a word or phrase. So, ontologies are a critical component for Artificial Intelligence (AI) and Natural Language Processing (NLP) applications.

Ontologies can be used to transform industry data into knowledge that can be used for several activities such as analysis on industry growth, improve information search. Ontology driven approach interprets knowledge in both human and machine-readable formats that enable knowledge exchange and sharing and enhances data integration, application interoperability and performance. The application of semantic technologies increases the flexibility of software solutions.

There are Big Data platforms running Hadoop, and its numerous components like MapReduce, Spark, Mahout, H2O Sparkling Water, and Kafka that enable us to access and analyse this data, but the fact remains that you must know the data before you can analyse it. Ontologies provide a mechanism for quickly ingesting data from Data Lake and making sense out of it. Data is available in heterogeneous formats such as spreadsheets, relational or NoSQL databases, semi-structured and unstructured documents. Integrating heterogeneous data sources and understanding data is complicated, and time-consuming task. Developing ontologies will help us to quickly understand a domain and accelerate the process.

This article demonstrates integrating heterogeneous data sources using domain specific ontologies. Further, the integrated data can be exposed as a knowledge base. Rest of the article is organised as follows - section 2 gives a brief state of the art on ontology driven approach at various application areas, section 3 illustrates knowledge interpretation from diverse data sources using domain ontologies and section 4 concludes.

2.Background Work

There is an increased usage of ontology-driven approaches to support requirements engineering activities, such as elicitation, analysis, specification, validation and management of requirements. Diego et al., [3] presented state of the art ontology driven approach in requirements engineering. In [4] Mike Bennett described Finan-

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

cial Industry Business Ontology (FIBO) that defines unambiguous shared meaning for financial industry concepts. As a conceptual model, FIBO can be used to develop semantic technology-based applications that may be used to carry out novel types of processing of data. Financial industry can leverage ontology with the development of Financial Industry Business Ontology (FIBO), but its adoption has been slow. One reason for this is that the power of ontologies is not well understood.

Industries like medical and healthcare collect data daily. Gabriela et al., [5] proposed ontology driven framework that can build a shared repository of surveys that can be used for data collection. The lack of a common or shared understanding of compliance management concepts is a barrier to effective compliance management practice. Norris et al., [6] developed ontology intended to provide a shared conceptualisation of the compliance management domain for various stakeholders.

Gozde et al., [7] proposed a delay analysis ontology that may facilitate development of databases, information sharing as well as retrieval for delay analysis within construction companies. Knowledge discovery and data mining (KDDM) approaches do not sufficiently address key elements of data mining model management (DMMM) such as model sharing, selection and reuse. Furthermore, they are mainly from a knowledge engineer's perspective, while the business requirements from business users are often lost. To bridge these semantic gaps, Yan et al., [8] proposed an ontology based DMMM approach for self-service model selection and knowledge discovery.

Claire et al., [9] proposed a reference ontology to support risk assessment for product-service systems applied to the domain of global production networks. The reference ontology helps accelerate the development of information systems by way of developing a common foundation to improve interoperability and the seamless exchange of information between systems and organisations. Malik et al., [10] proposed an ontology-based approach to address several key issues relating to interoperability for collaborative product development within networked organisations such as aerospace industry. The work includes semantic data mediation model to ensure interoperability, a cloud-based platform to enable complex collaboration scenarios.

Despite the growth of video game industry, there is a lack of interoperability that allows developers to interchange their information freely and to form stronger partnerships. Janne et al., [11] presented the Video Game Ontology (VGO), a model for enabling interoperability among video games and enhancing data analysis of gameplay information. Lepuschitz et al., [12] described an ontology-based knowledge model for the batch process domain to achieve a semantic data model for vertical and horizontal data integration. An important asset of this knowledge representation is the consideration of industrial standards and thus its applicability in conjunction with a commercial supervisory control and data acquisition system. El-Sappagh et al., [13] proposed and implemented a new semantically interpretable fuzzy rule-based systems framework for diabetes diagnosis. The framework uses multiple aspects of knowledge-fuzzy inference, ontology reasoning, and a fuzzy analytical hierarchy process (FAHP) to provide a more intuitive and accurate design.

Seth Earley [14] details how NASA used ontologies to define the data sets in terms that could be interpreted and enable integration with external data sets. It is being observed that ontologies are widely using in diverse domains such as healthcare, video game, aerospace, information technology etc. This article describes few cases where industry is dealing with heterogeneous data sources.

3.Knowledge Interpretation from Diverse Data Sources

Industries in various domains have data from diverse data sources such as spreadsheets (like csv, excel), relational databases (like MySQL, Oracle, PostgresSQL, H2 etc.), NoSQL databases (like Hadoop, Cassandra, MongoDB etc), web services and semi structured and unstructured documents (like ppt, doc, pdf etc.). Domain data may spread across these heterogeneous data sources used for different business functions. To perform comprehensive analytics, searching across the data sources and fetching right information from these data sources, is quite complex. The issues with the data sources are heterogeneous formats, syntax and semantics. Due to these issues, industries could not perform holistic analysis, share and exchange of data among various departments of an organisation. Ontology is a good remedy to address these kinds of issues. Ontology provides means to integrate heterogeneous data sources and bring the data sources into homogeneous syntax and semantics that enables information sharing, exchange in machine understandable format and application interoperability.

The basic components of the ontology are concepts, properties (data properties and object properties), instances and inferred data. The ontology elements can be broadly classified into two categories namely TBOX data and ABOX data [15]. RDFS facilitates the description of ontology elements with the help of knowledge representation data models in the form of classes/concepts and properties known as TBox. RDF is a standard data model that interchanges and merges instance data, known as ABox. TBox is associated with classes/concepts and ABox is associated with instances of the classes. Knowledge base is a combination of TBOX, ABOX and inferred data. The basic idea of interpreting knowledge from heterogeneous data sources is creating

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

instances for the ontology concepts from the data sources, describing the instances and linking the instances with meaningful relationships. Figure 1 shows a generic architecture of ontology driven approach.

This paper describes knowledge interpretation from 1) structured data sources and 2) unstructured data sources.

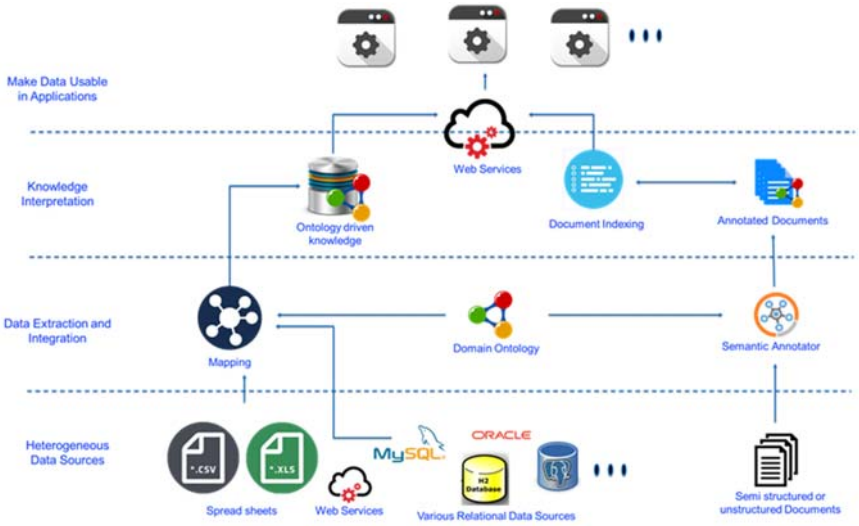


Figure 1: Architecture of Knowledge Interpretation from Diverse Data Sources

In both the cases, ontology development plays a crucial role. Developing an ontology includes [16]:

- Defining classes in ontology
- Arranging the classes in a taxonomic (subclass–superclass) hierarchy
- Defining slots and describing allowed values for these slots
- Filling in the values for slots for instances.

Determining the domain and scope of the ontology is another important activity of the ontology development process. Answers to the following questions will help to determine the domain and scope of the ontology.

- What is the domain that the ontology will cover?
- For what we are going to use the ontology?
- For what types of questions, the information in the ontology should provide answers?
- Who will use and maintain the ontology?

3.1. Hybrid Mathematical model for ontology concepts.

The mathematics beyond the ontology is more of approximations in the domain situation models and set of domain knowledge models .Enhanced mathematical model for the ontology problems with theory Concepts [17]

Let us consider OP is the representation of the logical relationship system with the approximation of a function UA(<OP,k>) with the domain situation k∈ En(OP) models. OX is the logical relationship system without quasi equivalent parameters, OP determines the approximation U A(< OX , k >) of the same set of intended situation models.were k ∈ En(O) is a knowledge model of the domain,∪ A(< O, k >) , where O is an enhanced logical relationship system k∈ En(O) that is a model of the ontology ,k ∈ En(O), then A(<O, k'>) ⊆ A(<O, k>) for any k' ∈ En(O).

Logical theorem for ontology

Equation 1:let h:En(OP) → k∈ En(OX) En(OX) is a mapping technique which eliminates a parameters with a enhanced logical relationship systems and H = {h(k) ☐ k ∈ En(OP)}. Then UA(<OX ,k >) = UA(<OX ,h(k)>) ∪ k∈ En(OX) k∈ En(OP)

Equation2: Let h : En(OP) → k∈ En(OX) En(OX) s a mapping technique which eliminates parameters with a enhanced logical relationship systems and H = {h(k) ☐ k ∈ En(OP)}. Then UA(<OX ,k >) =

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

UA(<OX ,h(k)>) ∪ k∈ En(OX) k∈ En(OP) ,U A(< OX , k >) but the theorem is enhanced with eliminating parameters and logical relationship systems k∈ En(OX)ÄH UA(<OX ,h(k)>) = UA(<OP,k >) , i.e. UA(<OX ,k >) = UA(<OP,k >) ∪ k∈ En(OP) k∈ En(OP) k∈ En(OX) k∈ En(OP) U A(< OX , k >) .The final approximation function is k∈ En(OX)ÄH system OX, is less precise than the approximation represented by the system OP. In the proposed mathematical model OP is the domain ontology which represented as a the logical relationship with the approximate function En(OP) of the set of the knowledge models , En(OX) of the same set of knowledge models. In the equation H is a subset of En(OX) is the approximation in the set of knowledge models OX also is less precise than the approximation determined by the system OP.Let us consider the equation 1 and 2 domain ontology model is a pure mixed unenriched logical relationship system OP = <Φ, P> with parameters.in the proposed equation k ∈ En(OP), then h(k) = Φ' ∪ Φ" the propositions belonging to Φ' are deduced from every proposition of Φ, and set of the ΦX ∪ Φ' ∪ Φ"is a semantically correct applied logical theory where ΦX is the set of all the propositions of Φ which contain no parameters, i.e. H ⊂ En(OX).The basic hybrid enhancement is in-terms of concepts, relations of the axioms, the axioms are linked with the every relations and concept and includes in OWL, RDF ontologies.This forms the conjunction of the maximum value of the each relation with diverse concepts in the domain of interest called (OX) U(MaX Fun(OP). For example Employee A works in department B and has financial accounting skills.Where Employee, department and skill and are concepts and works in ,has skill are relation. The axioms are expressed as constraints like “some Employee” and “only skill ”.

3.2.Knowledge Interpretation from Structured Data Sources

In case of structured data sources like spreadsheets, web services and databases, a knowledge base can be created by defining individual instances of the ontology classes. However, the integration of heterogeneous structured data sources is another challenge. Before creating knowledge base, creating virtual data layer that can be accessed through JDBC connectivity is a good idea. Once the virtual data layer is created, knowledge base creation can be done through mapping relational elements to ontology constructs. Ontology Based Data Access (OBDA) [18] technique enables conceptual view of a (relational) data source that abstracts details about data storage. This conceptual view is realised through an ontology that is connected to the data source through declarative mappings, and query answering. Further, the mappings and ontology files can be used to make knowledge available through web services or simple natural language (NLP) queries.

3.2.1 Knowledge Interpretation from Unstructured Data Sources

There are two different cases for defining individual instances of domain ontology classes from unstructured documents.

Case 1: Document metadata is available in a structured format
Case 2: Document metadata is not available in a structured format

3.2.2 Interpreting Knowledge from Unstructured Documents and Metadata

If metadata of the document is available, then it is a straightforward approach. Using the approach described in section 3.1, the metadata can be exposed as knowledge base, further the knowledge base can be used for annotating the documents. Annotation process links document content with ontology classes, named individuals. This enables semantic search on the unstructured documents through the knowledge based derived from metadata.

3.2.3 Interpreting Knowledge from Unstructured Documents with no Metadata

This approach includes NLP process and supervised learning apart from ontology. Initially, ontology is used to annotate document content with ontology concepts manually. The annotated content can be used for defining individual instances of the ontology classes. Further the manually annotated documents can be used to develop a supervised learning model. The supervised learning model helps to annotate named entities from the documents with respective ontology classes. The pair of entity and classes can be used to define instances and linking the instances with predefined semantic relationships.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

4.Conclusion

This paper demonstrates ontology driven knowledge interpretation from diverse industry data sources. Many industries like financial, healthcare, Information technology etc. are having heterogeneous data sources like relational databases, spreadsheets, unstructured documents etc. To perform holistic analysis or querying the data sources, there is a strong need of integrating these heterogeneous data sources. Ontology is a prominent semantic technology that enables creating domain ontologies, defining named individuals from heterogeneous data sources and relationships among them. Using ontology, we can derive both human and machine-readable knowledge base from the data sources. Since the knowledge is interpreted in machine understandable format, it can be easily shared and exchanged among various applications of the domain and used for improved natural language search to fetch any connected or related information from disparate data sources with ease.

References

1.T. R. Gruber, Toward principles for the design of ontologies used for knowledge sharing. In: International journal of human-computer studies. Kluwer Academic Publishers, Dordrecht, 1993, 907–928.
2.S. Randall, *Why Ontologies?*, 2017 available at https://www.datasciencecentral.com/profiles/blogs/why-ontologies
3.D. Diego, V. Jéssyka, I. B. Ig, C. Jaelson, I. Seiji, B. Patrick, S. Alan, *Applications of ontologies in requirements engineering: a systematic review of the literature*, Requirements Engineering, 21(4), 2016, 405-437.
4.B. Mike, *The financial industry business ontology: Best practice for big data*, Journal of Banking Regulation, 14(3-4), 2013, 255-268.
5.H. Gabriela, L. Laura, K. Daniel, H. Hlomani, S. Deborah, B. H. Philip, *An ontology-driven approach to mobile data collection applications for the healthcare industry*, Network Modeling Analysis in Health Informatics and Bioinformatics, 2(4), 2013, 213-223.
6.S. A. Norris, I. Marta, S. Shazia, *Compliance management ontology – a shared conceptualization for research and practice in compliance management*, Information Systems Frontiers, 18(5), 2016, 995-1020.
7.B. Gozde, D. Irem, B. M. Talat, *An ontology-based approach for delay analysis in construction*, KSCE Journal of Civil Engineering, 22(2), 2018, 384-398.
8.L. Yan, A. T. Manoj and O. B. Kweku-Muata, *Ontology-based data mining model management for self-service knowledge discovery*, Information Systems Frontiers, 19(4), 2017, 925-943.
9.P. Claire, N. U. Esmond, N. Ali, P. Dobrila, P. Keith, I. M. Y. Robert, *An ontology supported risk assessment approach for the intelligent configuration of supply networks*, Journal of Intelligent Manufacturing, 29(5), 2018, 1005-1030.
10.K. Malik, F. Nicolas, F. D. S. Catarina, G. Parisa, *A cloud-based platform to ensure interoperability in aerospace industry*, Journal of Intelligent Manufacturing, 27(1), 2016, 119-129.
11.P. Janne, R. Filip, G. Daniel, P. V. María, I. Jouni, P. Jari, G. Asunción, *An ontology for videogame interoperability*, 76(4), 2017, 4981–5000.
12.W. Lepuschitz, A. Lobato-Jimenez, A. Grün, T. Höbert and M. Merdan, *An industry-oriented ontology-based knowledge model for batch process automation*, 2018 IEEE International Conference on Industrial Technology (ICIT), Lyon, 2018, 1568-1573. doi: 10.1109/ICIT.2018.8352415
13.S. El-Sappagh, J. M. Alonso, F. Ali, A. Ali, J. H. Jang and K. S. Kwak, *An ontology-based interpretable fuzzy decision support system for diabetes diagnosis*, in IEEE Access. doi: 10.1109/ACCESS.2018.2852004
14.S. Earley, *Really, Really Big Data: NASA at the Forefront of Analytics*, in IT Professional, 18(1), 2016, 58-61. doi:10.1109/MITP.2016.10
15.Sunitha Abburu, Suresh Babu Golla, *Effective Partitioning and Multiple RDF Indexing for Database Triple Store*, Engineering Journal, 19(5), 2015, 139-154.
16.N. F. Noy and D. L. McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*, 2001.
17. Alexander Kleshchev, Irene Artemjeva, "MATHEMATICAL MODELS OF DOMAIN ONTOLOGIES", International Journal Information Theories & Applications, Vol 14, 2007
18.L. Davide, G. Xiao, D. Calvanese, *Cost-Driven Ontology-Based Data Access*, 2018, arXiv.1707.069974

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

THE MODELING AND SIMULATION OF WIRELESS CAMPUS NETWORK

Oyenike Mary Olanrewaju
Computer Science and Information Technology Department
Federal University Dutsinma, Katsina State, Nigeria
oolanrewaju@fudutsinma.edu.ng

ABSTRACT

In recent time wireless area network has come to stay as an indispensable tool of communication in education sector. It has foster dissemination of information among various stakeholders in education sector with minimal stress and justifiable cost. This research designed a Wireless Campus Local Area Network (WLAN) for Sa'adatu Rimi College of Education (SRCOE). The design was modeled using OPNET IT GURU Academic Edition. The analysis was carried out in scenarios to determine the performance of the network. The design interconnects the college computers in such a way that communication can flow from one department to the other. The network model includes connection to Internet. The results of the performance evaluation revealed minimal network delay for ftp applications and minimal response time for http requests. It also revealed that the designed can accommodate addition of client's computers to about 100% of the present number of computers without performance degradation.

Keywords: Local Area Network, Delay, Response time, Application, Education

1 INTRODUCTION

Wireless technology is the technology of delivering data from one point to another on a computer based network without using physical wires. Some of these wireless components are Access Points, Antenna, Routers, wireless bridges, servers and wireless workstations.

The present day networking environment has been taking over by wireless and sensor technologies such as IEEE 802.11 series. Mobile devices such as ipads, laptops and personal computers are wirelessly enabled either in IEEE 802.11a, b or g standard. The most recent standard of 802.11n comes in different forms. Wireless technology due to ease of implementation has been replacing the wired networks and gradually making networking an indispensible tool in Education system and other life endeavors. With Wireless networking, the cables are eliminated where wireless alternative can be provided leaving wired network for backbone network. Considering the various benefits of the wireless network, many tertiary institutions have provided Wireless Local Area Network (WLAN) on campus for students and faculties. The advantages of wireless networking are more obvious in tertiary institutes because of the dynamic environment and the ease of accessibility of information and records (Han, 2008)

This research designed a Wireless Campus Network for the Sa'adatu Rimi College of Education. A discrete event simulation package OPNET, which can compute the time that would be associated with real events in real-life situation was used. Software simulator is a valuable tool especially for today's network with complex architecture and topologies. Designers can test their new ideas and carry out performance related studies. It is therefore, free from the burden of trial and error of hardware implementations. The concern here is developing the overall campus network to meet critical design goals: efficient traffic flow, mobility, and scalability need of academic environment.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The rest of the paper is organized as follows: related research works are discussed in section 2, the methodology of the network design and analysis discussed in section 3 while section 4 and section 5 contains results and conclusions respectively.

2 REVIEW OF RELATED WORK

Zubairi and Zuber (2000) evaluated the network performance of State University of New York, Fredonia campus under usual network load conditions. Time-sensitive applications such as voice over IP were used for the network transmission. The research presented results for network behavior under typical and heavy load conditions showing some potential bottleneck points. Two client systems were configured for interactive voice located in two different subnets to pass the traffic through the core campus network backbone. The results show very good performance under typical load conditions. However, the delays and delay variations increase under loaded conditions. This reflects the low scalability of the network in the institution.

Han (2008) carried out a review of usage of WLAN from student's perspectives. This study presents the results of the research on the use of on-campus WLANs from two campuses of UNITEC, New Zealand. The data was gathered from literature review, observations surveys and interviews with students who were privileged to have access to WLANs at either campus. The research analyzed the data and outlined opportunities for education sectors to improve their current wireless networks.

In Siunduh (2013) data were collected from 14 Universities representing 21% of all Universities in Kenyan. In investigating into the growth involved in deployment of wireless campus network, security level implemented and how it affects internet usage in Kenyan Universities, stratified random sampling and purposeful sampling were adopted to come up with the sample size. The respondents were ICT Managers, network administrators, students, finance officers and researchers. The major findings of this study indicate annual increase of internet usage that ranges from 41% to 142% with wireless campus network deployments. For universities where wireless networks are not deployed, growth in internet usage ranges from 0-20%. Universities with more restrictive security level implementation attracts high number of users seeking support from IT team to access wireless networks and therefore limits internet usage on campus.

Schwab and Bunt (2004) analyzed the usage of campus-wide wireless network. A week-long traffic trace was collected, recording address and protocol information for every packet sent and received on the wireless network. A centralized authentication log was used to match packets with wireless access points. The trace was analyzed to answer questions about where, when, how much, and for what are wireless network was being used for. The College of Law among others achieved a high rate of wireless adoption and maximized the value of wireless networking for its students.

Mengdi Ji (2017) This paper introduced the theoretical part of WLANs, including advantages and limitations of WLANs, protocol standards of WLANs, components of WLANs, different topologies of WLANs and also the security of WLANs. The contents also included the basic process of planning a campus WLAN, designing a topology by using VISIO and planning APs by using Hive Manager NG and Floor planner tools. It also lists hardware and the budgets with the selected components. The research is good for studying the basic idea about WLAN design.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3 RESEARCH METHODOLOGY

Having reviewed related literatures, other research procedure includes preliminary investigation which was carried out among the academic staff of the College using a questionnaire and interview as tools. Deans of various schools were interviewed to determine the ICT facilities available in the college and the level of computer literacy as well as the internet surfing ability of the staff. The existing information flow was analyzed, the proposed WLAN designed, the network was modeled and configured in OPNET. Simulations and analysis of result performed for performance evaluation.

3.1 System Analysis

In this phase of the analysis process for the existing systems, the primary objective is to uncover the problem or opportunity that exists in a system without computer network. The summary of the information flow within each school is represented section 3.2. The College has five (5) schools and a total of twenty seven(27) departments, a library and an administrative unit which are all equipped with computer equipment but no computer network exists in the College as at the time of this research . Schools and the Number of Departments are as follows: Natural and Applied Sciences - Seven (7) Departments; School of Education Eight (8) Departments; Vocation - Four (4) Departments; Languages - Three (3) Departments; Arts and Social Sciences - Five (5) Departments.

3.2 Survey Findings

a) The information required from one school or department to another comprises of student records, continuous assessment report and other related academic reports.
b) From Table 3.1 about 59% of the staff surf internet for less than 3hours a week. This will definitely improve if wireless LAN and internet facilities were made available for staff use. E-books, journal, e-mails and lecture notes constitute the major focus of staff Internet desires
c) All academic staff laptops have Ms-word, Ms-excell, Ms-access, CorelDraw, Autocads, as application running on them.
d) All the staff laptops have wireless capabilities

Table 3.1: Department and the Number Staff Surfing hours per week

DEPARTMENT	Questionnaire. Distributed	Questionnaire Returned	No. of Staff weekly Surfing		
			<3hrs	<5hrs	>5hrs
Physics	9	7	1	6	0
Chemistry	11	10	1	7	2
Inter Science	11	8	4	4	0
Computer Science	12	10	3	4	3
Agric Educ	21	20	10	9	1
Home/Econs	26	22	16	4	2
English	18	16	12	3	1
History	16	14	5	8	1
Social Studies	12	11	4	4	3

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Arabic Medium.	30	28	26	2	0
Primary Educ.	25	20	15	5	0
General Studies	22	19	9	7	3
Fine Arts	30	28	15	10	3
Geography	23	22	12	8	2
Admin	39	33	26	5	2
Totals	305	268	158	86	23
Percentage		87.87%	59%	31.2%	8.6%

Summary of Information Flow from One Department to the Other

- Correspondence
- Student's exams result from one department to the other
- Journals
- Lecture materials
- Enquiries by staff from administrative department

3.3 System Design

The goal of this design is to provide interior mobile coverage for the various departments in the schools. This will allow the staff to access internet for their various academic activities.

3.3.1 General Design Requirements and Considerations

A campus LAN is expected to grow at a moderate rate from time to time. The scalable blueprint assumes that the campus LAN has the following design requirements and characteristics Terri and Haller (1998):

a. Medium to large network size (from 300 up to 15,000 nodes)
b. Centralized and distributed servers
c. Some bandwidth-intensive applications, including multimedia
d. Internet/wide-area connectivity
e. Minimum traffic priority needed
f. Security needed on the network
g. Moderate growth rate
h. Mobility occurrences (implemented with wireless technologies)

3.3.2 Design Considerations

Each AP and antenna combination produces a single area of coverage. Each of these single areas is referred to as a cell. Multiple overlapping cells are used to provide wireless coverage for areas larger than a single cell alone can produce. According to Quellet et al (2002), the distance of each node from an access point affects the data rate. The data rate decreases as the coverage area increases until there is no coverage at all.

To position wireless equipment for optimal network performance, Mitchell (1999) recommends the following guidelines

1. First and foremost, *do not settle prematurely on a location* for the wireless access point or router, experiment; try placing the device in several different promising locations. While trial-

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and-error may not be the most scientific way to find a good spot for your equipment, it is often the only practical way to assure the best possible performance.

2 Consider installing the wireless access point or router in a *central location.*

3 Next, *avoid physical obstructions* whenever possible. Any barriers along the "line of sight" between client and base station will degrade a radio signal. Plaster or brick walls tend to have the most negative impact on network transmission.

4. *Avoid reflective surfaces* whenever possible. Signals literally bounce off of windows, mirrors, metal file cabinets and stainless steel countertops, lessening both network range and performance.

5. Likewise, *install the unit away from electrical equipment* that also generates interference. Avoid electric fans, other motors, and fluorescent lighting.

6. If the best location one can find is only marginally acceptable, consider adjusting the base station antennas to improve performance. Antennas on wireless access points and routers can usually be rotated or otherwise re-pointed to "fine tune" signaling.

7. Follow the specific manufacturer's recommendations for best results.

3.3.3 Design Map

Design Map is a representation of each department according to their location within the college. The buildings were represented as blocks in figure 3.1 in order to imagine how the network signals can flow across the campus. Point to multipoint topology was used in all the design. Figure 3.1 is representing a situation where the router for the design is placed in the central library and all the other departments receive the transmission from there through their access points. The design was modeled using OPNET and various simulation scenarios were carried out to obtain the best possible performance evaluation result.

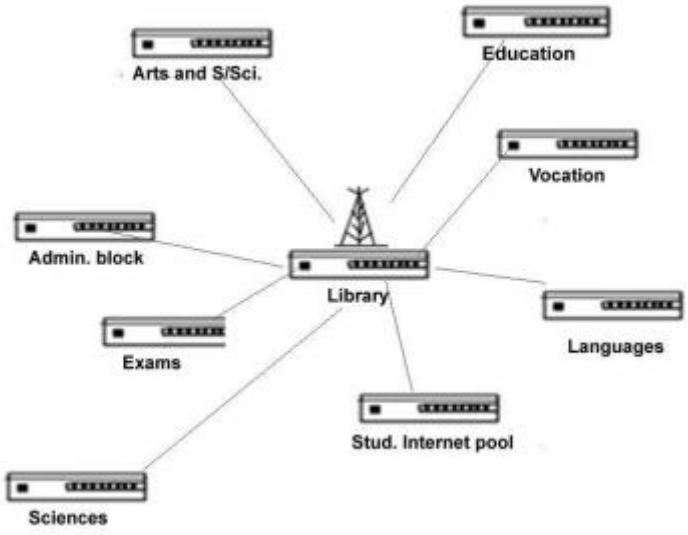


Figure 3.1: Design Map for the proposed WLAN for Sa'adatu Rimi College of Education

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3.3.4 Expected Focus for the Proposed Network Design

The proposed design is expected to meet the communication demand of members of staff and students within the school and from one school to the other.

1. Communication within the Schools
Each school in the college comprises of various departments. The subnet of each school is modeled with access points, required number of computers and other peripherals. The intra-communication within each school is established wirelessly through the access point and a server in the Examination office and Administrative block. This represent a basic service set topology.

2. *Communication between various schools*
This foundation is built on extended service set. The basic service set of each school communicates with each other through the wireless access points. The backbone of the network is modeled with internet router and omnidirectional antenna either at the library or computer complex as scenarios. The model is built with file transfer protocol (ftp) server and one hypertext transfer protocol (http) server. The http server serves web applications while the ftp server serves all file requests from the nodes.

3.4 Modeling the WLAN

- Network subnet were created in OPNET for various schools and departments
- Each school subnet has access point for intra-communication within the school computers.
- The statistics were configured in each scenario as required to determine the respective Network performance
- Various Simulations were ran to represent various network user demands
- The results were analyzed and presented in graphs and figures.

3.4.1 The Network Scenario
In this scenario, the college has eight (8) major subnets with each subnet having an access point for communication within the subnet and communication with other access points. The Http server, ftp server, wired router and the antenna is placed at computer science complex. The consideration for this is to allow the server to be at close proximity to computer lecturers for management purpose. It has a total of six hundred and two (602) work stations. Ethernet-4-switch was used to connect the router and the access point. Figure 3.2 and 3.3 represent the backbone network and computer complex subnet respectively.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

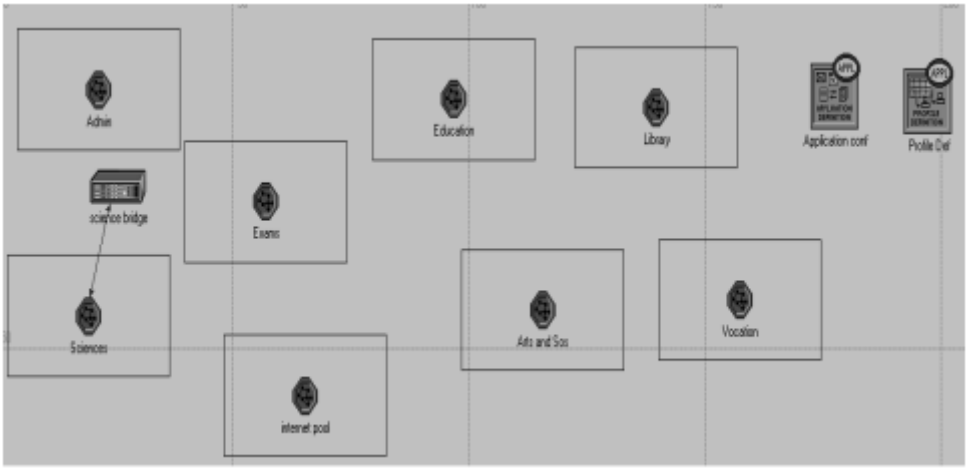


Figure 3.2: SRCOE WLAN

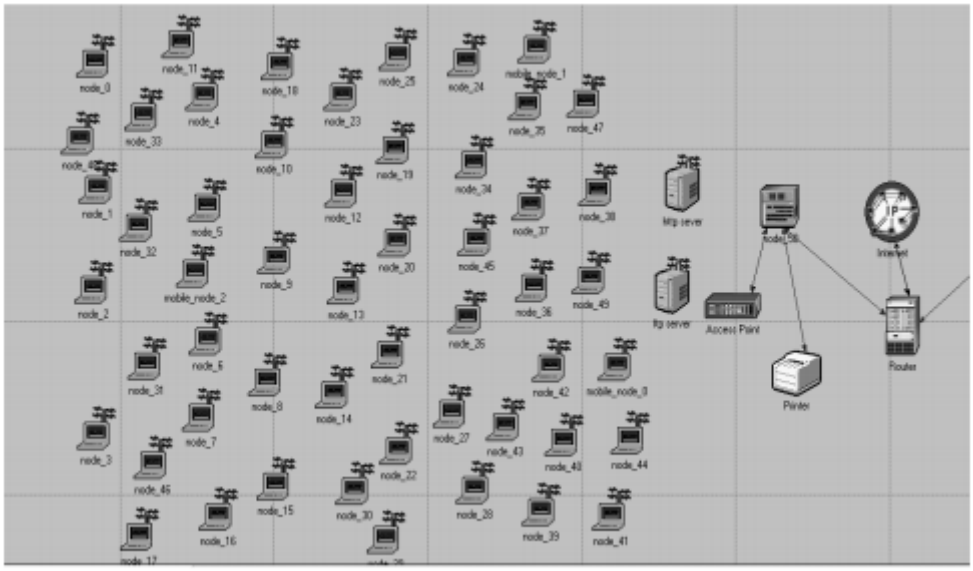


Figure 3.3: Computer Complex Subnet

Table 3.2 Scenario Configurations

Nodes	Applications	Simulation Statistics
Http Server	Web browsing, Email	Time: 8hrs real life
FTP server	File transfer(heavy)	Period of simulation: 7min 19 secs
Admin server	Application supported services- database heavy	Data rate: 11Mbps
Router- Ethernet2-slip8-gateway		

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

602 wireless work stations	ftp, http and database applications	
Ethernet4_switch		

4. RESULTS AND DISCUSSIONS

4.1 Results

The performance evaluation results were presented in figures and in graphical representation. HTTP page response time, Ftp download response time, Wireless LAN delay and Database Query response time were the focused performance evaluation parameters.

4.1.1 Http Page response time:

The time taken for a web page to be downloaded at the clients request measured in seconds. The Http response time from the simulation is reported in Table 4.1 and the graph is displayed in Figure 4.1 with max response time of 0.64sec

Table 4.1 Top Objects Report: Client Http. Page Response Time (seconds)

Rank	Object Name	Min	Average	Maximum	STD Dev
1	node_31 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0385	0.0531	0.124	0.0238
2	node_40 <Web user / Web Browsing (Heavy HTTP1.1)	0.0388	0.0496	0.108	0.0156
3	node_15 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0419	0.0480	0.074	0.0075
4	node_19 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0416	0.047	0.053	0.0036
5	node_20 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0403	0.047	0.057	0.0045
6	node_29 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0419	0.047	0.054	0.0048
7	node_47 <Web user / Web Browsing (Heavy HTTP1.1)	0.0424	0.047	0.064	0.0066
8	node_38 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0420	0.0475	0.054	0.0038
9	node_26 <Web user / Web Browsing (Heavy HTTP1.1)>	0.0415	0.047	0.057	0.0055

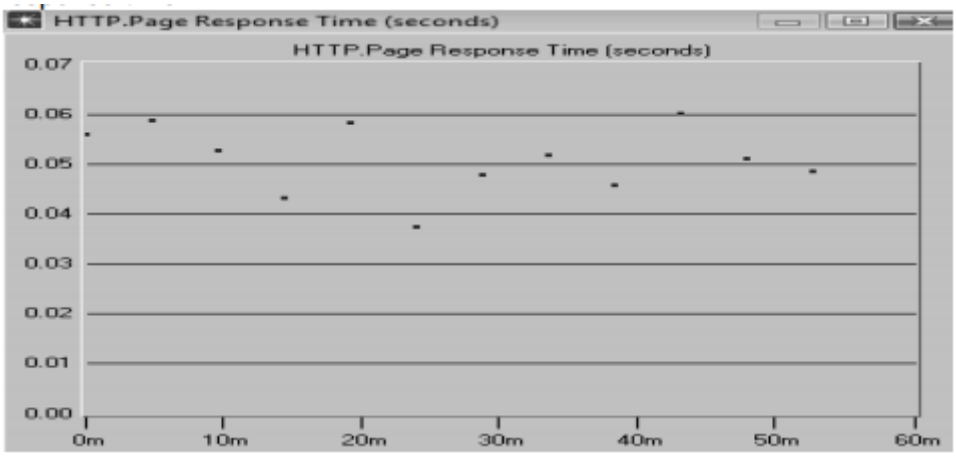


Figure 4.1 display of Http response time in second

4.1.2 Ftp Download

This is the measure of time taken for a file to be downloaded at the user's request. This is measured in seconds as represented in Table 4.2 and plot of the response time displayed in Figure 4.2

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 4.2 Top Objects Report: Client Ftp Download Response Time (Sec.)

Rank	Object Name	Mini	Average	Maxi	Std Dev
1	node_5 <file transfer / File Transfer (Heavy)>	0.390	0.390	0.390	0.000
2	node_32 <file transfer / File Transfer (Heavy)>	0.370	0.377	0.385	0.007
3	node 15 <file transfer / File Transfer (Heavy)>	0.330	0.370	0.408	0.032
4	node 16 <file transfer / File Transfer (Heavy)>	0.285	0.359	0.454	0.070
5	node_30 <file transfer / File Transfer (Heavy)>	0.229	0.356	0.456	0.066
6	node_46 <file transfer / File Transfer (Heavy) >	0.328	0.347	0.367	0.019
7	node 1 <file transfer / File Transfer (Heavy)>	0.321	0.344	0.371	0.021
8	node 17 <file transfer / File Transfer (Heavy)>	0.313	0.340	0.369	0.023
9	node_44 <file transfer / File Transfer (Heavy)>	0.220	0.339	0.398	0.065
10	node_45 <file transfer / File Transfer (Heavy)>	0.275	0.336	0.387	0.044

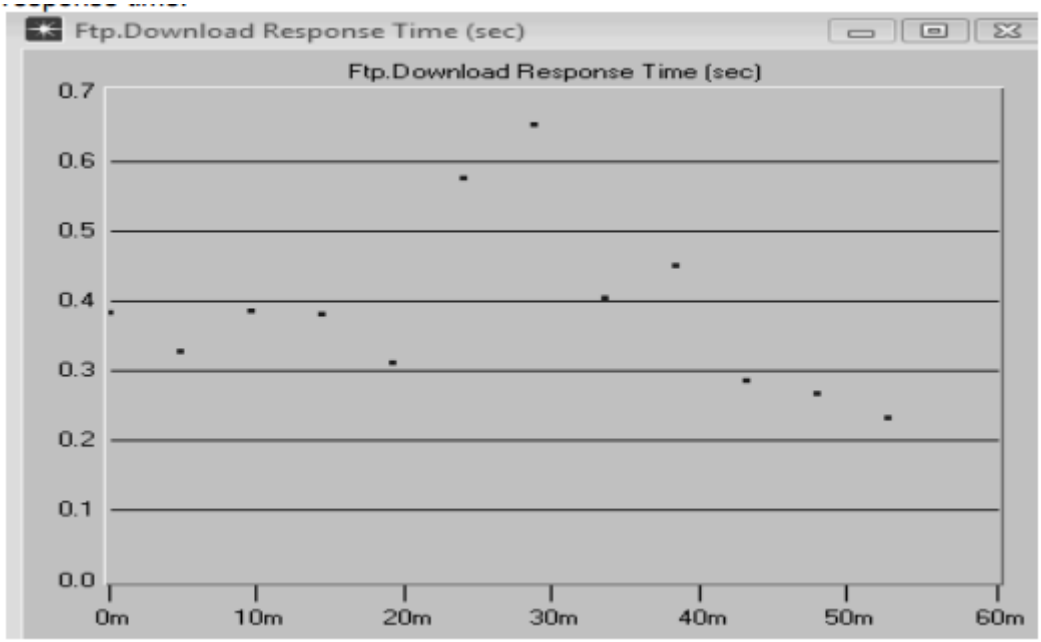


Figure 4.2: Client Ftp Download Response Time (Sec.)

4.1.3 WLAN delay

This is the general network delay experienced by the object nodes while on the network. This is measured in seconds. The minimum delay, maximum delay and average delay experienced by each node is represented in Table 4.3 and the graph represented in Figure 4.3.

Table 4.3 Top Objects Report: Wireless LAN. Delay (sec)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Rank	Object Name	Minimum	Average	Maximum	Std Dev
1	node_16	0.00204	0.00800	0.0206	0.00499
2	node_30	0.00254	0.00739	0.0180	0.00406
3	node_8	0.00406	0.00728	0.0103	0.00255
4	node_31	0.00256	0.00673	0.0116	0.00275
5	node_7	0.00218	0.00660	0.0105	0.00265
6	node_19	0.00249	0.00653	0.0096	0.00228
7	node_46	0.00207	0.00650	0.0157	0.00325
8	node_1	0.00217	0.00650	0.0104	0.00271
9	researcher	0.00247	0.00644	0.0177	0.00378
10	node_26	0.00239	0.00641	0.0122	0.00284

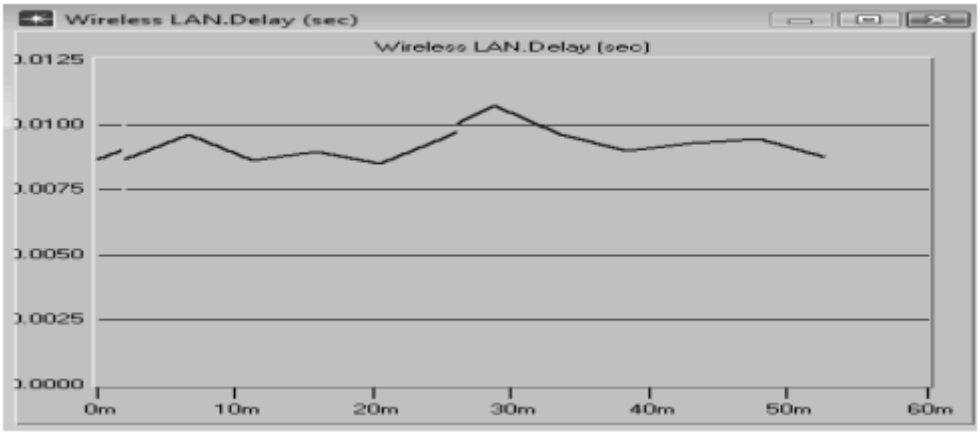
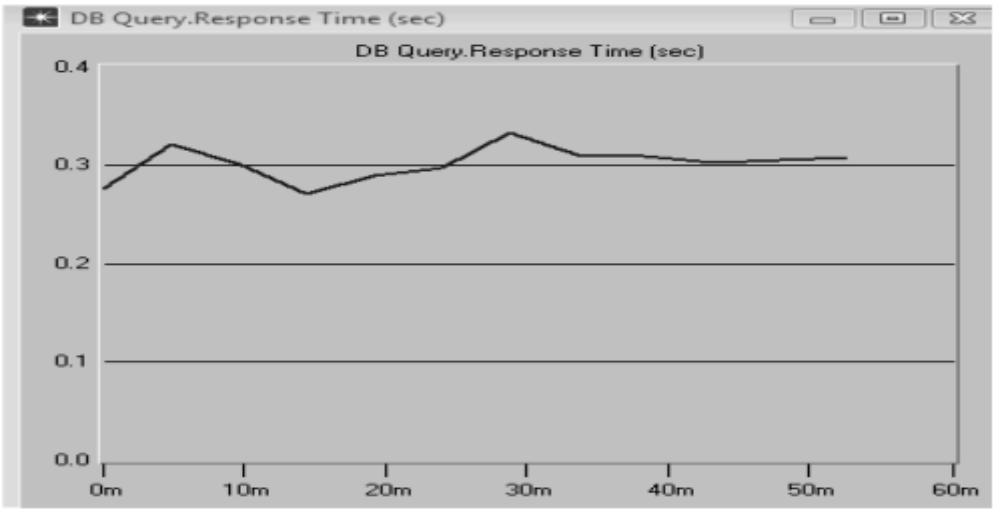


Figure 4.3: Wireless LAN. Delay (sec)

4.1.4 Database Query Response Time

The time taking for the node to receive response to database query across the network, it is measured in seconds. The graph in figure 4.4 was obtained from simulation in respect of Database Query Response time.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

4.2 Discussion

In a network environment, the load of applications running on the network and the capacity of the network devices are some of the major determinant factors of the values of network evaluations parameters. When the network was fully loaded the performance of the network is expected to vary from the statistics obtained when the network was lightly loaded. Wireless LAN delay obtained ranges from 0.01 – 0.02 sec. Ftp page response time has a maximum value of 0.065sec. HTTP page response Time has a max of 0.064sec.

5. Conclusions

The following conclusions were drawn from the research:

a) The performance evaluation revealed an Ftp down load response range of 0.0123 - 0.65sec which represent a good performance.

b) Http down load response time range of 0.007 – 0.064 also confirm that the network will give acceptable performance if implemented.

c) Wireless LAN delay range of 0.01 – 0.02 indicate a network that is fast enough for implementation.

REFERENCES

Mitchell B. (2008). How to Optimally Position a Wireless Access Point or Router, in: Your Guide to Wireless Networking, About.com

Ouellet E., Padjen R. and Pfund A. (2002). Building a Cisco Wireless LAN. (Rockland: Syngress Publishing, Inc)

Terri Q, .and Haller K. (1998) Designing Campus Networks, (Macmillan Computer Publishing)

Zubairi J. A. and Zuber M. (2000). State University of New York Fredonia Campus Network Simulation and Performance Analysis Using OPNET. Department of Mathematics and Computer Science, zubairi@cs.fredonia.edu, zube7270@fredonia.edu

Han Y. (2008) A thesis submitted in partial fulfilment of the requirements of the degree of Master of Computing UNITEC New Zealand 7 August 2008

Mengdi Ji M. (2017) Designing and Planning a Campus Wireless Local Area Network South-Eastern Finland university of Applied Sciences.

Siunduh E. S (2013) Analysis Of The Effect Of Wireless Campus Networks On Internet Usage In Kenyan Universities, School of Computing and Informatics, University of Nairobi

Schwab D. and Bunt R. (2004) Characterizing The Use Of A Campus Wireless Network Department of Computer Science University of Saskatchewan Saskatoon, SK S7N 5A9 Canada {das515, bunt}@cs.usask.ca

Hutchins R.and Zegura E. W. Measurements From a Campus Wireless Network, College of Computing Georgia Institute of Technology Atlanta, GA 30332-0280 fron,ewzg @cc.gatech.edu

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Relaxed Context Search over Multiple Structured and Semi-structured Institutional Data

Ahmed Elsayed
Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt
eng_ahmedyakoup@yahoo.com

Ahmed Sharaf Eldin
Prof. and Dean, Faculty of Information Technology and Computer Science, Sinai University; Faculty of Computers and Information, Helwan University, Cairo, Egypt

Doaa S. Elzanfaly
Informatics and Computer Science, British University in Egypt; Faculty of Computers and Information, Helwan University, Cairo, Egypt

Sherif Kholeif
Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt

ABSTRACT

The NoSQL graph data model has been widely employed as a unified relationship centric modeling method for various types of data. Such graph modeled data is typically queried by means of queries expressed in native graph query languages. However, the lack of familiarity that users have with the formal query languages and the structured of the underlying data has called for relaxed search methods (like, keyword search). Many research efforts have studied the problem of keyword search but in a single database setting. Relaxed query answering becomes more challenging when querying multiple heterogeneous data sources. This paper presents a technique for relaxed query processing over multiple graph-modeled data that represents heterogeneous structured and semi-structured data sources. The proposed technique supports various forms of relaxed search (including, context keyword, phrase, proximity search). An extensive experimental evaluation on a real world dataset demonstrates that the proposed technique is more effective than the existing keyword search methods in terms of precision, recall and f-measure.

KEYWORDS

Top-k query processing, Graph databases, Heterogeneous data, Relaxed search

1 INTRODUCTION

The increased accumulation of data has affected the way in which the data is modeled and queried. For example, many NoSQL data models have been proposed to respond to various modern application needs [1, 2]. One of these needs is the call for a unified representation of interconnected and complex data. Furthermore, assessing how data is correlated is essential for many tasks including exploring and querying the data. NoSQL graph data model has been widely employed to provide a unified, relationship centric modeling for complex and various forms of data (including, structured and semi-structured data) [3, 4]. Many research efforts have studied the problem of searching graph modeled data in a single database setting [5 - 11]. Numerous search methods have been proposed ranging from graph query languages (like, Cypher, SPARQL, .., etc.) to flexible query answering (like, keyword search). Although graph modeled data is typically queried by means of queries expressed in native graph query languages, these query languages require users to have a good knowledge of both the query language syntax and the structure of the underlying data. Consequently, systems will be hard to use especially for non-technical users. On the other hand, keyword queries are easy to formulate but they lack the structure support. As a result, keyword queries are ambiguous and usually produce large numbers of answers many of which are irrelevant. Therefore, users usually have to reformulate their queries multiple times before they can get the desired information.

The problem becomes more complicated and challenging when querying multiple heterogeneous data sources. As an example, institutional data and applications usually encompasses data from diverse sources. The data is often large, heterogeneous and dynamic in nature. In such scenarios, users need to be able to search multiple forms of data through simple search interface.

In this paper we present a unified technique, called ConteSaG, for transparently querying multiple heterogeneous data sources with various types of relaxed queries. A comprehensive survey of related techniques is presented in our previous paper[12]. The proposed technique models the structured and semi-structured data in different sources as data graphs where nodes represents entities and edges represents relationships between them. Then, a context search is enabled as a query model over these graph representations. Specifically, users are allowed to express their information needs in various forms of relaxed queries (including, keyword, phrase, proximity queries) with simple context specification exploiting the partial knowledge they have about the entities and their attributes. In that way, the proposed query model improves the expressiveness of the keyword-based approach and in the same time does not require to precisely expressing information needs in forms of formal query language.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 briefly introduces the necessary preliminaries needed to understand the proposed technique. The proposed technique is described in section 4. Section 5 demonstrates experimental evaluation. Finally, section 6 concludes the paper and sketches out future work.

2 RELATED WORK

A large body of research work has studied the problem of flexible query answering (specially, keyword search) over a single data source. For example, many systems have been proposed in the literature to apply keyword search on relational databases[13, 14, 7, 8, 9, 15, 16, 17, 18, 19]. Another line of work applies keyword search in the context of XML repositories [20, 21, 22]. Some other techniques have been proposed to enable keyword search over RDF data [23, 24] and data streams [25]. The keyword query processing employed by the above methods and systems can be broadly categorized into two main approached; schema-based approach and data graph-based approach. Given a set of keywords, the schema-based approach exploits schema graph of the underlying data source to create candidate queries and then execute the generated queries to obtain the results. On the other hand, the second approach represents the data and their relationships in a data graph where nodes represent entities and edges represent relationships between them. At the query time, results are generated by traversing the data graph to find trees or sub-graphs that connect nodes containing query keywords.

Apart from the previous approach, another line of work addresses the problem of keyword search as a translation problem. Representative examples include [26], [27] and [28] that transforms keyword queries into SPARQL queries, Xpath queries and formal queries, respectively. Authors in [29] have proposed a method that interprets keyword query to generate an interpretation tree whose nodes corresponds to decisions. Each decision in turns represents a question to the user that needed to be answered to construct the final query. They utilized Ontologies to reduce number of questions. A template search approach for exploring relational databases has been proposed in [30]. Specifically, it generates a set of natural language question on the fly that may represent the search interpretation of the keyword query. The generated query patterns are then ranked and translated into SQL statements that are in turns executed to retrieve query results. Similarly, a technique in [31] interprets the keyword query to infer user intension. All possible interpretations are expressed in forms of query patterns. These patterns are then ranked and described in natural language for the user to select from.

However, all the above research efforts focus on querying a single data source and do not consider searching scenarios that require searching multiple data sources.

The proliferation of distributed data in several applications calls for a flexible way to query multiple heterogeneous data sources. In the literature, there are two main approaches to do this; namely, physical integration and virtual integration. The physical integration retrieves and integrates all the underlying data from various sources in a single centralized location and then performs the search over the integrated data [32, 33, 34]. Ease [32] proposed keyword search over index representation of graph modeled data that are extracted from multiple structured, semi-structured and unstructured data sources. At a preprocessing stage, the data is extracted from all the participating sources to be indexed and maintained in a centralized location. The index representation is exploited at the query time to find sub-graphs that contains query keywords. FleQsy [35] has introduced a framework for flexible query answering over graph modeled data. The authors have applied the framework to solve various problems including keyword search over relational databases [33], keyword search over RDF data [36] and approximate graph matching on RDF triple stores[37]. However, all these techniques do not support distributed implementation except the second one. Authors in [34] have proposed an approach for integrating structured and unstructured institutional data. The approach integrates diverse data models from multiple data sources in a unified RDF graph. A declarative query language (called in star) has been introduced to enable formal search over that unified graph data. However, using formal query language to explore and search data affects the usability of the proposed method negatively as users have to master the syntax of the query language in order to be able to express their information needs. Moreover, the techniques belong to this approach are hard to implement in many practical scenarios. This is because there are always several management and security issues that make it hard to integrate all data from various sources in one single central location.

The second approach, virtual integration, to query multiple data sources is to connect the underlying sources by providing a unified view and way to query the underlying data [38 - 43]. In [38], a solution called Kite has been proposed to address the problem of keyword search over heterogeneous relational databases. Kite belongs to schema-based approach in that it exploits the schema of the data sources to model them as schema graphs. At query time, it receives user queries in form of set of keywords and generates set of candidate networks (CNs) over the schema graphs of the data sources. Finally, the generated CNs are translated into SQL queries to be executed in the corresponding databases. Unlike Kite that is designed to work over a small number of database (up to ten as the authors stated), Q system [39] enables keyword search over a large number of databases(web scale scenarios).Q scans the meta data of the data sources to create a search graph that composes the relations, their attributes and the relationships between them. The search graph is expanded at query processing time by adding keyword node for each keyword in the query. Nodes are then linked to all graph nodes whose labels match the keywords. The resulting query graph is traversed to generate minimal joining trees that may produce relevant answers. Finally, SQL queries are generated and executed in the local sources. Q is extended in [44] to automatically incorporate new sources and it is further extended in [45] to actively soliciting feedback for query answers. Authors in [40] have proposed Keymantic framework for keyword based search over multiple relational databases. Keymantic addresses

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the keyword search as a translation problem. Specifically, it utilizes intentional knowledge about data sources (e.g., DB schema, semantic knowledge and rules specified by users) to transforms keyword queries into SQL queries. The translation process goes through two main steps. First, keyword query is analyzed to discover the intended meaning. The second step generates the interpretations, in forms of SQL queries, that best describe the intended semantic of the query. Apart from keyword query relaxation, authors in [41] have presented another form of query relaxation. In particular, the authors have proposed a technique that extends the existing SQL rewriting techniques to work in cases where there is incomplete partial mapping between sources.

While all the above methods consider searching over relational data, the work in [42] presented a technique for top-k query processing over multiple heterogeneous XML data sources. The technique accepts XML queries in form of tree pattern and leverage the schema graph of the underlying XML repositories to produce relaxed queries. The evaluation of the resulting queries over the corresponding XML repositories is scheduled using a threshold strategy.

Unlike the methods outlined above that work on a specific type of data, a method discussed in [43] combines a concept-based mediator system with schema-based keyword query processing to enable keyword search over structured and semi-structured data in a single system. The system employs YACOB concept based mediator system to describe the data in the underlying sources in form of concept schema. When receiving keyword queries, the system interprets the query to generate set of concept based queries. Those queries are optimized and translated to local queries to be executed on the corresponding sources. However, as this technique belongs to schema-based approach, it inherits its limitations. For example, it may generate a large number of candidate networks many of them do not produce any answer when execution. Moreover, this technique depends on single word IR index that leads to many irrelevant answers in case of phrase search or when query concepts composes more than one word.

The proposed technique in this paper differs from the systems belong to the physical integration approach in that it does not need to integrate all data in a single central location but instead it represents data in local sources as graphs and let that graph representations distributed. The employed generic graph representation of the data allows the proposed technique to work over various forms of data including structured and semi-structured data. Furthermore, the proposed system does not need to create complex mediated schema and translate the global queries against this schema to large numbers of local queries, as in the case of the virtual approach systems. Instead, the proposed technique enables various types of relaxed context queries (including, keyword, phrase, proximity queries) over the graph representations of local sources. The proposed query formulation allows users to contextualize the search terms by determining the context in which the search terms occurs. In that way, many irrelevant results produced by the existing keyword search methods could be excluded.

3 PRELIMINARIES

In this section, we formulate some concepts related to modeling heterogeneous data, the specification of queries, and the answer semantic.

Unified Data Model. The proposed technique models various types of heterogeneous data (including, structured relational databases RDB, semi structured xml repositories and RDF stores) as data graphs. The adopted data model is generic in a way that represents data as a connected graph G(V,E) where V is a set of nodes representing real world entities and E is a set of edges that represents the relationships between them. For example, in case of a data graph representing a RDB, V corresponds to tuples and E represents foreign key references among that tuples. Whereas, V represents XML elements and E represents parent-child (or, id-idref) relationships in case of XML documents.

Query Specification. Users are allowed to specify context queries. A context query basically consists of a set of terms. Each term is associated with a context(called contexted term). Formally, the query specification can be defined as follows.

$$Q: = C_i{:}Term_i; \ldots ; C_m{:}Term_m$$

Where, a term can be one of three types: (1) single keyword (2) sequence of keywords (3) a phrase. The second type allow user to perform proximity search by specify a value for Pdistance parameter. Pdistance determines the maximum distance in which two consecutive keywords should appear. For example, if the user wants to search for papers that contain the term “graph database” in its title. The query is formed as: paper.title: graph database.

Answer Semantic. Given an m context query $\{C_1T_1,\ldots,C_mT_m\}$ against a set of data graphs $\{G1(V,E),\ldots, Gn(V,E)\}$ representing n data sources. The answer to such query is top-k distinct rooted trees. A distinct rooted tree is a tree containing at least one matched node for each query term (in case of AND semantic) or matched nodes for some of the query terms (in case of OR semantic). A node matches a query term $C_i.T_i$ if it contains the term T_i in the corresponding context C_i. The matched nodes are the leaves of the answer tree and there is at most one answer tree rooted at each node; hence the tree called distinct rooted tree.

4 SYSTEM ARCHITECTURE

The system architecture is shown in Figure1. The user interface allows users to issue context queries Q_{glob} (called, global queries). The interface displays the answer trees as well. Generally, a global query is transformed to a set of local queries Q_{targ} (called, target queries). These queries are then submitted to the corresponding targeted sources to be processed. Finally, the query results are then combined and re-ranked to form the final answers to the global query. Specifically, the query processing is performed over two layers: **meta-search layer** and **local search layer**. The meta-search layer performs a lightweight schema mapping offline. At the query time, it processes queries through five sub components: query parser, data source selector, query

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

rewriter, query disseminator and results re-ranker. The functions of these components are explained in the below subsection.

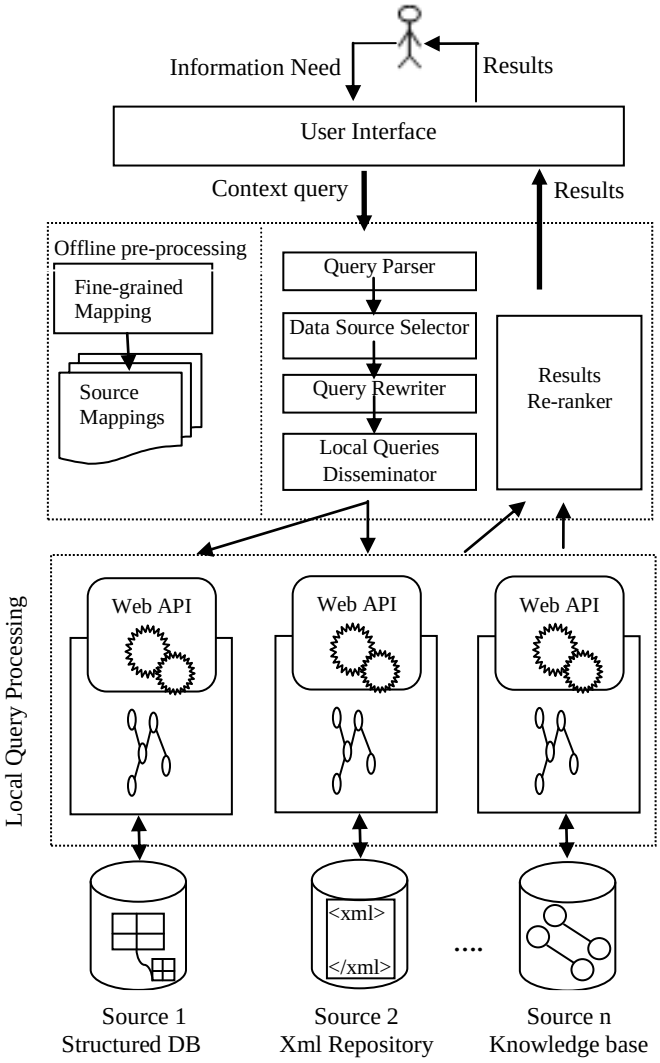


Figure 1: ConteSaG Architecture

4.1 Meta Search Layer

Fine-grained Mapping. Basically, each data source consists of a set of entities. Each entity corresponds to a real-world object described by a set of attributes. Therefore, this component uses schema matching tools to offline computes same-as-mapping $M_s$for the underlining entities and attributes of the local sources. In particular, the mapping process is all about aligning entities and attributes taking into consideration the attributes' values and data types. Global names are then given to the mapped entities and attributes. For instance, let there are two entities named Proceeding and Article in two different data sources and both of them contain an attribute called Title. If the Proceedings.Title is mapped to Article.Title, the global name for these mapped contexts can be any of their local context names (e.g., Article.Title) or it can be left to the system administrator to give the global name (e.g., Paper.Title).

Query Parser. The main role of the query parser is to pinpoint entities and attributes involved in the global query, given that each term in the query is annotated with a context.

Data Source Selector. This component identifies the data sources that cover **ALL** the searched constructs (i.e., entities and attributes identified by the Query Parser) in case of AND semantic or **some** of them in case of OR semantic using the source mapping M_s. In other words, for each distinct context in the global query, the Selector identifies the data sources that contains the corresponding entity and attribute and create a list of them called term source list. The generated lists are then intersected to find out the sources that cover the maximal possible number of query terms. The Data Source Selector is orthogonal to the existing data source selection method [48] that summarizes each database in a keyword relationship graph and then leverages that graphs to compute similarity between the user query and each database.

Query Rewriter. The Query Rewriter translates the global query Q_{glob} into a set of target queries $\{Q_{targ1, \ldots,} Q_{targn}\}$ to be executed on the selected data sources. Specifically, the query rewriter utilizes mapping M_s in hand to rewrite the global query that is posed over global constructs (entities and attributes) to new target queries that can be evaluated over the constructs of the selected data sources.

Query Disseminator. The target queries are then disseminated to the selected sources via calling the web APIs of these sources to be processed in parallel. The details of how the target queries are processed locally is presented in section 4.2.

Results Re-ranker. As a post processing step, this component receives the answer trees from local sources and re-rank and combine them to find the final top-k answer trees. The re-ranking process can take into account many scoring components but without loss of generality currently the re-ranker uses the original tree score LScore(T) as it comes from local sources and penalty the trees that do not cover all search terms. The below equation is used to recalculate the score of sources trees.

$$Score(T) = LScore(T) - \left(\frac{NumberOfMissingTerms}{NumberOfQueryTerms}\right) * \lambda \quad (1)$$

Where, LScore(T) is the original local tree score and λ is the answer tree compactness weight.

4.2 Local Search Layer

Offline Preprocessing. Generally, a common approach to enhance online performance is to perform some offline computation. Therefore, in order to effectively enable context search over heterogeneous data sources, ConteSaG constructs two important structures for each data source: Context-aware Positional Index (*CaPI*) and Node Neighborhood Index (*NNI*). *CaPI* is used to efficiently map keywords appearing in the attribute values to their entities. Furthermore, the *CaPI* maintains the positions of the indexed keywords in the attribute value of the corresponding entity instance. Formally, a *CaPI* entry consists of a key and Positional Posting List PPL <*key, PPL*>. The *CaPI key*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

composed of two parts; a keyword followed by a context (keyword >Context). While, the *PPL* is a list of {*id; PL*}; where *id* is an instance id and *PL* is a list of positions that maintains where the keyword appears. As an example, let the keyword *W* appears in a value of the attribute *A* of entity *E*. Consequently, there should be an entry/row in *CaPI* with key *W>E>A* (assuming > is a special character reserved only for indexing purposes and any other delimiter that never occurring in the indexed keywords and contexts works as well). For each instance *I* of entity *E* containing *W* in attribute *A*, there is a post in the *PPL* as follows $Id_I;<p_1,p_2, \ldots>$. One way to reduce the key size and consequently the *CaPI* size is to encode the context. For example, the key *Ontology>Article>Title* can be stored as *Ontology>1>1*. All encoding configuration of all entities and their attributes are maintained in a context encoding file.

As mentioned above, the second structure that is built for each data source at the preprocessing stage is the Node Neighborhood Index (*NNI*). *NNI* maps each content node *u* in the data graph representing a data source to its neighborhood. The neighborhood of a content node *u* is a list of triple *<v,w,p>*, where *v* is a node connected to *u* via a shortest path *p* of weight *w*. Clearly, *p* is the nodes constructing the shortest path from *u* to *v*. In order to restrict the exploration of neighbors, the neighborhood only maintains the nodes that are connected to u with paths of length *Dmax* or less. Neighbor nodes in each neighborhood are sorted ascending by path weight.

By maintaining *NNI*, there is no need to store the whole data graph in memory. Instead, only the neighborhoods of the query matched nodes are loaded and hence overcome the scalability issues related to working with very large graphs.

Online Processing. Given a context query Q_{targ} against a data source or equivalently the corresponding data graph G(V,E), the local search layer generates and reports the top-k answer trees as a result to the query. More specifically, when a web API of a local source got invoked, the query is processed in four coarse steps as shown in algorithm 1. First, the content/ matched nodes are located for each term in the query. To do this, the query is splitted to find the contexted terms (line 2). Then, the algorithm decomposes each contexted term to its two parts: the context and the term (line 3 and 4). As contexts are encoded in *CaPI*, line 5 gets the encodes for query contexts by the help of context encoding file. Recall that the query term can be a sequence of keywords, complete phrase or a single keyword. Thus, line 6 tokenizes the term to extract words and remove stop words (if any). After that, *CaPI* is looked up to find content nodes that contain the query term (represented by set of words) in the given context (represented by encoded context). Finally, if a query term consists of multiple words, then the resulting posting lists of these words are intersected (with taking Pdistance into account) to find the content nodes for this query term (line 8).

The algorithm in the second step (line 10 and 11) looks up the *NNI* to load the term neighborhoods for each query term. The value of *Udmax* is set by the user to restrict the maximum path length between any content node and the root of its answer tree. Therefore, *Udmax* is used to load the neighbors that are connected to content nodes with path of length *Udmax* or less. Step three creates a term Cursor C_i for each term neighborhood constructed in the previous step. The created cursors are initialized and exploited to traverse the corresponding term neighborhoods in increasing weight order. That means the nodes with smallest path weights are retrieved first. Finally, in step four, the initialized term cursors are called in around robin manner to construct top-k answer trees (algorithm 2).

Algorithm 1: Source Local Search (Contxt Query **Q**, Max proximity distance **Pdistance** :=1, Threshold **Udmax**, Context-aware Positional Index **CaPI,** NodeNeighborhood Index**NNI,** Context Encoding File)

```
//Step1: Find content nodes for each context term in query
1: ContentNodesPerTerm List := ∅/* a list to hold a list of
matching nodes for each query term*/
2: QueryContextTerms [ ] ← Query.Split(';');
3: Foreach ContextTerm∈ QueryContextTerms do
4:     ContextAndWords [ ] ← ContextTerm.Split(':');
5:     EncodedContext ← getEncodeForContext (
                ContextAndWords [0], ContextEncodingFile);
6:     Words[ ]← tokenize (ContextAndWords [1])
7:     PositionalPostingLists PPLs ← Look up CaPI (
                EncodedContext, Words);
8:     TermContentNodes [ ] ← Intersect PPLs(PPLs, Pdistance);
9:     ContentNodesPerTerm .Add(TermContentNodes);
/* Step2: Look up neighborhood Index to load term
        Neighborhoods*/
10: Foreach TermContentNodes in ContentNodesPerTerm do
11:     neighborhood_ti← Look up neighborhood index NNI (
                TermContentNodes, Udmax);
  //Step3:Create cursor for each neighborhood
12: Foreach i ∈ [1,m] do        // Assume m-term query
      C_i← CreateCursor(neighborhood_ti) ;
 //Step4: Generate top-k answer trees
13: Construct the top-k distinct rooted trees by calling the created
    term cursors in a round robin manner (Algorithm 2)
```

Given term Cursors $C_i,\ldots,C_m$ and the size of result set K, algorithm 2 computes the top-k distinct rooted trees as answer trees. In order to efficiently compute the answer trees incrementally, the algorithm utilizes the principles of the threshold algorithm (TA) proposed in [47].

Generally, the algorithm access nodes iteratively from term neighborhoods to compute the k root nodes with the best structural compactness scores. More specifically, each root node r corresponds to a root node of a candidate answer tree. The compactness score of an answer tree rooted at node r is computed as the sum of its path weights connecting its root to its leaves (that represent content nodes). Clearly, the scoring function is a cost function. Therefore, the trees with the best scores are those with the smallest sum of path weights.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Algorithm 2: Generate top-k Answer trees

Input: Set of term cursors **{C_i, …,C_m},** Required number of answers **K**

Output: Top-k Answer trees

Initialization: Create an empty candidate dictionary $D_{cand} := \Phi$**,** Create an empty priority queue to hold results $Q_{Wk} := \Phi$**,** Create an empty list to maintain the ended cursors $Ended_C := \Phi$

// Growing Phase

1: **While** (t>T & $Ended_C$. Count< m)
2: C_i := getNextCursor();
3: r := C_i . next();
4: **if** r == NULL **then**
5: $Ended_C$. Add(C_i);
6: **else**
7: **If** r has not been seen from C_i before **and** r is already seen from all ended cursors $Ended_C$ **then**
8: Update D_{cand} (r, i, w_i, p_i);
9: **If** r is now accessed from all cursors **and** score(r) <t **and** there is no isomorphic tree in $Q_{Wk}$to the current tree **then**
10: UpdateQ_{Wk} to include r;
11: Update T;

// Preparing for Shrinking Phase

12: Construct Virtual Lattice $\mathcal{G}$;

// Shrinking Phase

13: **While** (t > u & $Ended_C$. Count< m) **do** */*u is the smallest upper poundγ_x^{ub} (∀ r in D and r has been accessed from all ended cursors and r ∉ Q_{Wk}) */*
14: C_i := getNextCursor();
15: r := C_i . next();
16: **if** r == NULL**then**
17: $Ended_C$. Add(C_i);
18: **else**
19: **if** r ∈ D_{cand} **and** r is already seen from all ended cursors **and** r has not been seen from C_i before **then**
20: Update D_{cand} (r, i, w_i, p_i);
21: v_r^{prev} := lattice node where r belonged;
22: **if** r was leader in v_r^{prev} **then**
23: update leader for v_r^{prev} ;
24: **if** r is now accessed from all sources **and** score(r) <t **and** there is no isomorphic tree in Q_{Wk} to the current tree **then**
25: update Q_{Wk} to include r;
26: **else if** r not accessed from all Cursors yet **then**
27: check if r is the leader of the new node v_r := v_r^{prev} ∪C_i;
28: $u := \min\{\gamma_v^{ub} : v \in \mathcal{G}\}$; *// use lattice leaders*
29: Construct and output all result trees in Q_{Wk} as top-k answers;

Essentially, Algorithm 2 operates in two phases: a growing phase and a shrinking phase. Starting with the growing phase, the term cursors are called in a round robin manner by invoking getNextCursor() (line 2). Then, the selected cursor C_i advances by calling Next() method that returns the next node *r* (line 3). If the cursor is not ended and therefore returns a node then, the growing phase in line 7 checks if the returned node *r* was not accessed from C_i before and r is already has been accessed from all ended cursors $ended_C$ (in case of AND semantic). If so, the algorithm maintains *r* in the candidate dictionary D_{cand} to indicate that r is retrieved from cursor C_i with a path P_i of weight w_i (line 8). After updating candidate dictionary, the algorithm in line 9 checks if r can be added to the result Q_{wk}. A Q_{wk} is a priority queue that preserves the set of k root nodes with the best scores. A root node r is added to Q_{wk} if the following three conditions hold. (1) *r* is now accessed from all cursors. That means r is a root for a tree that has a matching node for each query term. (2) the score of r is better than t. score t is the k^{th} best score in Q_{wk}. (3) there is no isomorphic tree t_j in Q_{wk} to the current tree t_i rooted at *r*. Two trees t_i and t_j are considered isomorphic to each other in two cases. The first one occurs when there is one-to-one mapping from nodes of t_i to nodes of t_j. The second case happens if the current generated tree rooted at *r* has one child for the root *r*. Such tree is also pruned as the tree formed by removing the root r would be generated and would be better answer.

At the end of each iteration of the growing phase, the algorithm updates *T* (line 11). The value *T* defines a threshold for the aggregate score of candidate roots that never retrieved from any cursor yet. Formally, T = $Sum(l_1, \ldots, l_m)$ where l_i is the last weight retrieved from cursor C_i. The growing phase terminates when *t* becomes less that or equals *T* or when all cursors are ended.

Whenever *t* is less that or equals *T*, this means that there is no root node which has not retrieved from any cursor yet can end up in the top-k results. Consequently, the algorithm progresses to the shrinking phase. During this phase the algorithm maintains the upper pounds γ_r^{ub} for the set of candidate root nodes that can end up in top k results. To do so efficiently, the algorithm constructs a virtual lattice $\mathcal{G}$ (line 12). The lattice $\mathcal{G}$ contains a node for every combination υ in the power set of Cursors $C_i,\ldots,C_m$. Each node υ in $\mathcal{G}$ maintains the ID of its leader r^{υ}, which is the node with the best partial aggregate score seen only in υ.

The shrinking phase iteratively calls the not-ended cursors in a round robin manner (line 14) and terminates when *t* is not larger than the smallest leader upper pound score. At each iteration, the shrinking phase checks whether the retrieved node *r* is already presents in the candidate dictionary D_{cand} or not. As no new node can end up in the top *k* results, the retrieved node should be presented in D_{cand}. If $r \in D_{cand}$, r is already retrieved from all ended cursors (in case of AND semantic) and r has not been accessed from this cursor C_i before (line 19), then the algorithm proceeds as follows: First, it updates D_{cand} to indicate that r is now retrieved from cursor C_i with a path P_i of weight w_i (line 20). Second, the algorithm in line 22 checks if *r* was the leader of the lattice node υ_r^{prev} where *r* belongs to before it was accessed from the current cursor C_i. If this is the case, a new leader for the lattice node $\upsilon_x^{prev}$is selected (line 23). After that, the algorithm checks whether *r* can be added to Q_{wk}. To do so, the algorithm checks the seam three conditions of adding a candidate root node during the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

growing phase (line 24). If all the three conditions holed, the result queue Q_{wk} is updated to include r. Otherwise, the shrinking phase checks if r is not accessed from all cursors yet (line 26). If so, r is promoted from υ_x^{prev} to the parent node in $\mathcal{G}$ that contains C_i besides the other cursors where r has been seen. Finally, the algorithm checks whether r becomes the leader there (line 27).

After the shrinking phase terminates, the algorithm constructs and computes the overall score of answer trees for all root nodes in Q_{wk} and returns them as top-k answer trees. The overall scores of the resulting answer trees take into account both the structural compactness of the answer tree Cscore(T) and the degree of content matching of the content nodes in the answer trees to the query terms Mscore(T). More specifically, the score of an answer tree T is the combined weighted sum of Cscore(T) and Mscore(T) and can be defined as follows.

$$Score(T) = \lambda * Cscore(T) + (1-\lambda)Mscore(T) \quad (2)$$

Where, λ is a weight used to control the relative importance between the two scoring components; Cscore(T) and Mscore(T). The compactness score Cscore(T) is based on the tree weight W(T) which is computed as the sum of path lengths from the root of T to its content nodes. The Cscore(T) can be defined as follows.

$$Cscore(T) = 1 - W(T)/MaxTreeWeight \quad (3)$$

The denominator is the maximum possible tree weight for the query that can be computed as (number of query terms * Udmax).The other scoring component Mscore(T) can be calculated as below.

$$Mscore(T) = 1 - 1/\sum_1^m Tf \quad (4)$$

Where, Tf is the term frequency of query terms in the content nodes of the answer tree T.

5 EXPERIMENTAL EVALUATION

The proposed algorithms for ConteSaG are implemented using visual studio 2013(Asp.net and C#). All experiments are conducted on AMD E1-1200 APU machine equipped with 350 GB SATA disk and 3 GB of memory.

Datasets and Queries. Due to the unavailability of agreed-upon benchmark for flexible search over multiple graph-modelled datasets, we follow a methodology similar to [48]. Specifically, we decompose DBLP [46] into five distinct subsets according to bibliography types. These five datasets are constructed as follows: D1 contains In-proceedings and is stored in a relational database (RDB). D2 maintains articles data and is stored in a RDB, as well. While D3, D4 and D5 contain master thesis, PhD thesis and book data, respectively. All these three sets are stored in XML documents. The statistics for these five data sets are shown in table 1.

Table 1: the DBLP subsets statistics

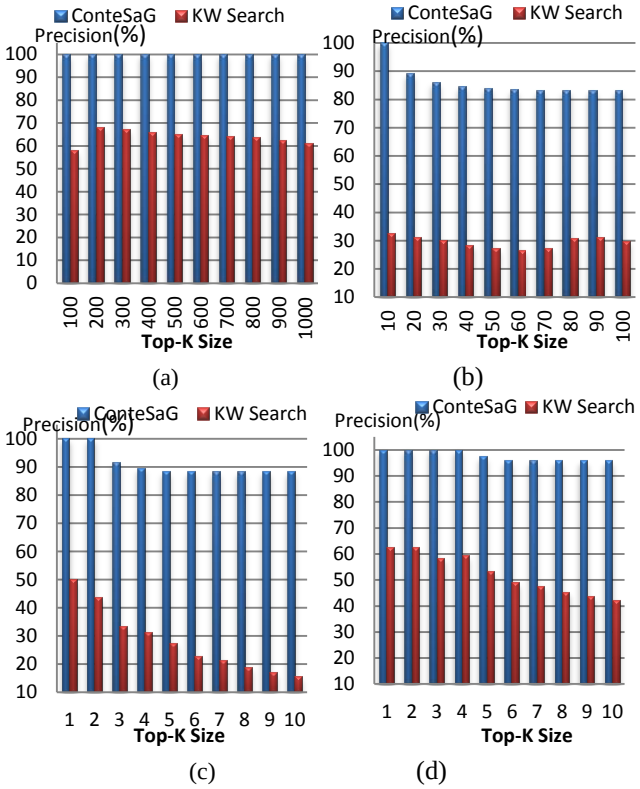
Dataset	Description	Format	# Nodes	# Edges
D1	In-proceedings	RDB	540902	1414904
D2	Articles	RDB	428691	999046
D3	PhD Thesis	XML	62375	123928
D4	Master Thesis	XML	29	40
D5	Books	XML	28271	44473

We created a test query set containing 45 different queries with different number of keywords and searched contexts (i.e., number of distinct entities and attributes needed to be explored). Furthermore, we take into account the query term occurrence in different data sets and the frequency of the query term in each set.

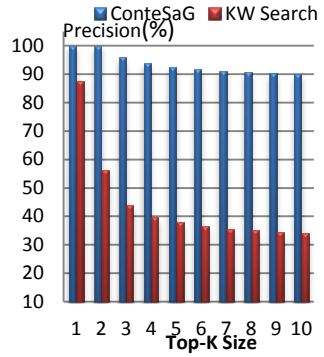
Effectiveness Evaluation. We employed three common effectiveness metrics: precision, recall and f-measure to evaluate the effectiveness of ConteSaG. We also compared ConteSaG to the existing keyword techniques like the one in [32]. Figure 2 and Figure 3 evaluate the precision and recall of both ConteSaG and the keyword search techniques as a function of query length L and max path length Udmax. The experiments involve four parameters: the result size K, max path length Udmax, the answer tree compactness weight λ and the proximity distance between consecutive keywords Pdistance. Table 2 summarizes the default and range of values of these parameters. For each experiment, a number of nine different queries with the same length but differ in number of searched contexts and the frequency of query terms are executed and the average precision and recall are calculated.

Table 2: Query Parameters

Parameter	Default value	Range of values
K	10	10-1000
Udmax	3	0–3
λ	0.7	0.7, 0.8
Pdistance	1	1 – 5



https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(e)

Figure 2: Precision Evaluation Result

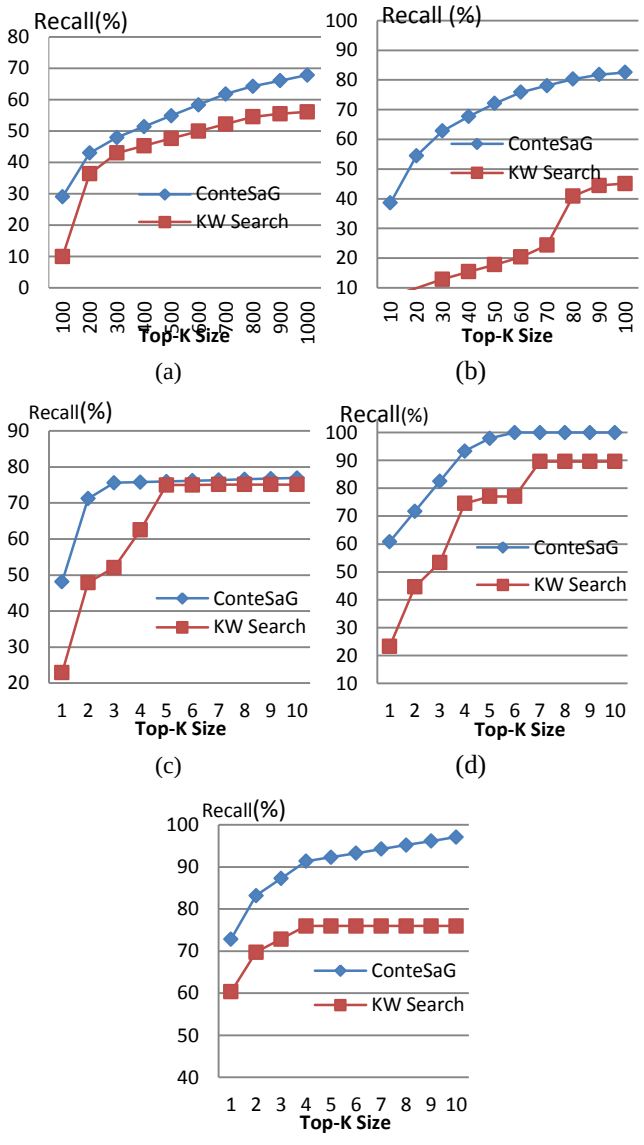
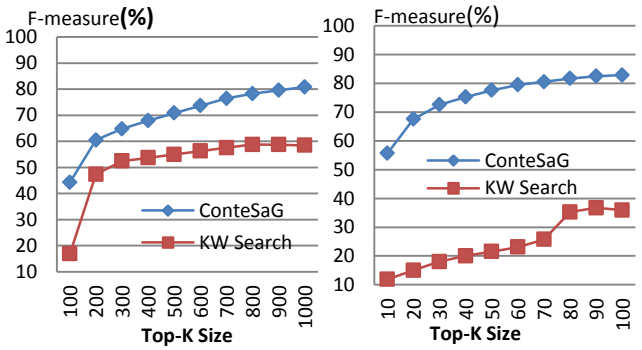


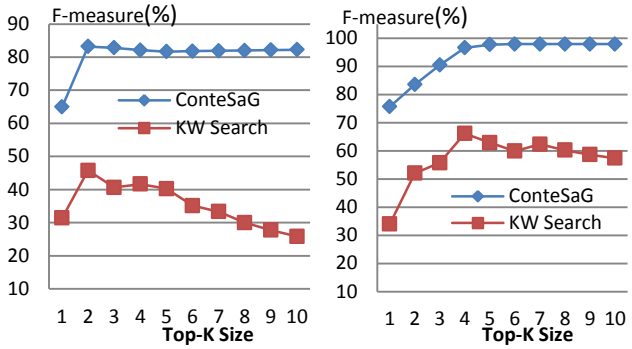
Figure 3: Recall Evaluation

Figure 2(a) to Figure 2(b) shows the precision rates as the number of keywords in query increases from 1 to 5 respectively. The results show that for large values of k, the precision of keyword search techniques drops with a larger percentage compared to ConteSaG. This can be attributed to two reasons: First, the keyword techniques depend on a traditional single-word inverted index to map keywords to the content nodes. Consequently, exploiting such indexes to retrieve the matching nodes that contain a sequence of consecutive keywords or for phrase search leads to a large number of false positive answers. On the other hand, ConteSaG depends on CAPI indexes that maintain positional information of the indexed keywords. The second reason is the lack of structure information (i.e., context of query terms) in keyword queries that leads to great ambiguity and result in many irrelevant data. While ConteSaG utilizes the context annotation of query terms to reduce the search space and finds the answer trees that best satisfy the query and hence the user information need.

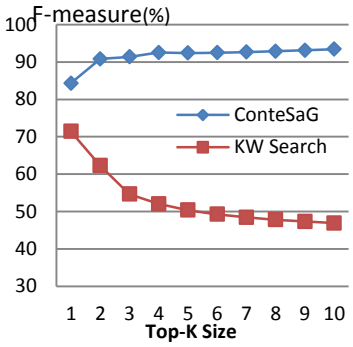
Figures 3(a - e) plots the recall curves over queries of size length ranging from 1 to 5 keywords. Overall, the recall steadily increases with k size. As expected, more returned results are likely to produce more relevant results. However, as Figure 3 illustrates, ConteSaG beats the other keyword search techniques and achieves better recall. This is because keyword search techniques in contrast to ConteSaG rank some irrelevant results higher than the other relevant results. Specifically, keyword search techniques for many queries return some compact answer trees (i.e., answer trees with small size) from irrelevant data sources and rank them higher than many other relevant answer trees of larger size. Furthermore, the same thing happens with the results that even come from relevant source. Contrary, this issue does not happen in ConteSag results. This is because at the beginning of the query processing the data source selector component selects only the relevant sources to be searched. Moreover, at those selected sources only the relevant entities and attributes participate and construct the search space to be explored.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(c) (d)



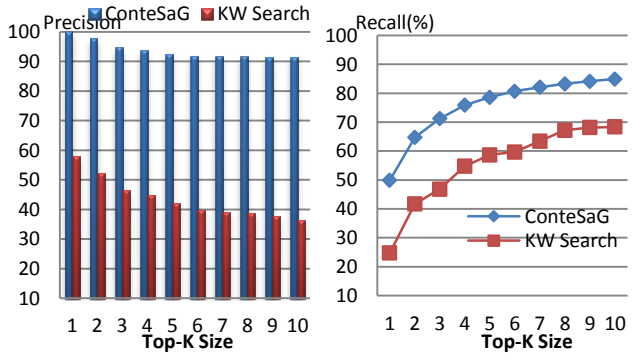
(e)

Figure 4: F-measure Evaluation

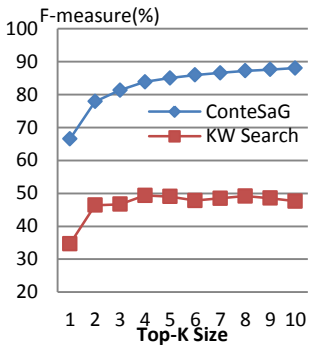
F-measure is used to evaluate the overall effectiveness of ConteSaG. F-measure is a harmonic mean of precision and recall and is computed as follows.

$$\text{F} - \text{measure} = \frac{2(P * R)}{P + R}$$

As shown in Figure 4, again ConteSaG outperforms keyword search techniques in terms of f-measure. This is logically as ConteSaG beats the other techniques in the two building measures; namely, precision and recall.



(a) Overall Precision (b) Overall Recall



(c) Overall F-measure

Figure 5: Overall Effectiveness Evaluation

Finally, the average of precision, recall and f-measure over 45 queries are computed and the results are shown in Figures 5(a), (b) and (c), respectively. The results show that ConteSaG is more effective than keyword search techniques. Specifically, ConteSaG improves precision, recall and f-measure by as much as 54%, 25% and 40%, respectively.

6 Conclusions and Future Work

This paper presents a technique that transparently enables context search over multiple heterogeneous graph modeled data. The proposed method works with various types of data including the structured and semi-structured data. Many forms of relaxed queries are supported. This includes context keyword, phrase and proximity queries. The experimental evaluation on a real world dataset demonstrates that the proposed technique is more effective than the existing keyword search techniques in terms of precision, recall and f-measure by as much as 54%, 25% and 40%, respectively.

In the future work we plan to extend our technique to support value heterogeneity besides schema heterogeneity. We also plan to improve search quality by clustering similar results and generating result snippets.

REFERENCES

[1] F. Gessert, W. Wingerath, S. Friedrich, and N. Ritter. NoSQL database systems: a survey and decision guidance. Computer Science-Research and Development, pages 353-365, (2017).

[2] A. Haseeb, and G. Pattun. A review on NoSQL: Applications and challenges. International Journal of Advanced Research in Computer Science, 8(1), (2017).

[3] Mike Buerli. The Current State of Graph Databases. (2012).

[4] J. Pokorny. Modelling of Graph Databases. Journal of Advanced Engineering and Computation, 1(1), pages 4-17, (2017).

[5] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault , S. Plantikow, M. Rydberg, P. Selmer, and A. Taylor. Cypher: An evolving query language for property graphs. In Proceedings of the 2018 International Conference on Management of Data, ACM, pages 1433-1445, (2018).

[6] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. W3C, URL: http://www.w3.org/TR/rdf-sparql-query/ (Document Status Update, 26 March 2013), (last visit: 2018).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan: Keyword searching and browsing in databases using BANKS. In Proc. 18th Int. Conf. on Data Engineering, pages 431–440, (2002).

[8] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar: Bidirectional expansion for keyword search on graph databases. In VLDB, pages 505-516, (2005).

[9] H. He, H. Wang, J. Yang, and P. Yu: BLINKS: Ranked keyword searches on graphs. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pages 305-316, (2007).

[10] L. Qin, J. X. Yu, L. Chang, and Y. Tao.: Querying communities in relational databases. In Proc. 25th Int. Conf. on Data Engineering, pages 724–735, (2009).

[11] D. Wang, L. Zou, W. Pan and D. Zhao. Keyword Graph: Answering Keyword Search over Large Graphs. In Springer, pages 635-649, (2012).

[12] A. Elsayed, A.S. Eldin, D.S. Elzanfaly, and S. Kholeif. ConteSaG: Context-based Keyword Search over Multiple Heterogeneous Graph-modeled Data. In Proceedings of Seventh ACM International Conference on Web Intelligence, Mining and Semantics (WIMS'17), Amantea, Italy, 6 pages. DOI: 10.1145/3102254.3102278, (2017).

[13] C. Aparna, and S.Bangar: Review on KeywordSeaech over Relational Databases. International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 11, (2014).

[14] M. Kargar, A. An, N. Cercone, P.Godfrey,JJ.Szlichta, and X. Yu. MeanKS: Meaningful keyword search in relational databases with complex schema. In SIGMOD, (2014).

[15] Jaehui Park and Sang goo Lee. Keyword search in relational databases. Knowl. Inf. Syst., 26(2), pages 175–193, (2011).

[16] Li, Guoliang; Feng, Jianhua; Zhou, Xiaofang; Wang, Jianyong: Providing built-in keyword search capabilities in RDBMS. In VLDB, pages 1-19, (2011).

[17] S. Agrawal, S. Chaudhuri, and G. Das: DBXplorer: A System for keyword Search over Relational Databases. In ICDE, pages 5-16, (2002).

[18] V. Hristidis and Y. Papakonstantinou: DISCOVER: Keyword search in relational databases. In Proc. 28th Int. Conf. On Very Large Data Bases, pages 670–681, (2002).

[19] V. Hristidis, L. Gravano, Y. Papakonstantinou: Efficient IR-Style Keyword Search over Relational Databases. In VLDB, pages 850-861, (2003).

[20] G. Li J. Wang, L. Zhou. Efficient keyword search for valuable LCAs over xml documents. In proceedings of the 16th ACM Conference on information and knowledge management, pages 31-40, (2007).

[21] Y. Xu, and Y. Papakonstantinou. Efficient LCA based keyword search in XML data. In Alfons Kemper, Patrick Valduriez, NoureddineMouaddib, Jens Teubner, MokraneBouzeghoub, Volker Markl, Laurent Amsaleg, and IoanaManolescu, editors, EDBT 2008, 11th International Conference on Extending Database Technology, Nantes, France, March 25-29, 2008, Proceedings, volume 261 of ACM International Conference Proceeding Series, pages 535–546, (2008).

[22] J. Li, C. Liu, R. Zhou and W. Wang. Top-k keyword search over probabilistic xml data. In Data Engineering (ICDE), 2011 IEEE 27th International Conference, pages 673-684, (2011).

[23] R. Virgilio, A. Maccioni, and P. Cappellari. A linear and monotonic strategy to keyword search over rdf data. In ICWE, pages 338-353, (2013).

[24] S. Han, L. Zou, J.X. Yu, and D. Zhao. Keyword Search on RDF Graphs-A Query Graph Assembly Approach. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 227-236, (2017).

[25] A. Markowetz, Y. Yang, and D. Papadias. Keyword search on relational data streams. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pages 605-616, (2007).

[26] S. Shekarpour, S. Auer, A. Ngomo, D. Gerber, S. Hellmann, and C. Stadler. Keyword-driven sparql query generation leveraging background knowledge. In web Intelligence, pages 203-210, (2011).

[27] D. Petkova, W. Croft, and Y. Diao. Refining keyword queries for xml retrieval by combining content and structure. In ECIR, pages 662-669, (2009).

[28] T. Tran, H. Wang, S. Rudolph, and P. Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped (RDF) data. In ICDE, pages 405-416, (2009).

[29] E. Demidova, x. Zhou, and W. Nejdl. Efficient query construction for large scale data. In SIGIR '13 Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pages 573-582, (2013).

[30] A. Zouzies, M. Vlachos, and V. Hristidis.Templated Search over Relational Databases. In CIKM, pages 21-30, (2014).

[31] Z. Zeng, Z. Bao, M. Li Lee, and T. Wang Ling. Towards an interactive keyword search over relational databases. In Proceedings of the 24th International Conference on World Wide Web, pages 259-262, (2015).

[32] Guoliang Li, Beng Chin Ooi, JianhuaFeng, JianyongWang, and Lizhu Zhou. EASE: an effective3-in-1 keyword search method for unstructured, semi-structured and structured data. In Proc.2008 ACM SIGMOD Int. Conf. On Management of Data, pages 903–914, (2008).

[33] R. De Virgilio, A. Maccioni, and R. Torlone. Graph driven exploration of relational databases for efficient keyword search. In GraphQ, pages 208- 2015, (2014).

[34] Luiz Gomes Jr. and Andr´eSantanch`. The Web Within: Leveraging Web Standards and Graph Analysis to Enable Application-Level Integration of Institutional Data. Springer-Verlag Berlin Heidelberg 2015, pages 26–54, (2015).

[35] A. Maccioni. Flexible query answering over graph-modeled data. In SIGMOD, pages 27-32, (2015).

[36] R. Virgilio and A. Maccioni. Distributed keyword search over RDF via mapreduce. In ESWC, pages 208-223, (2014).

[37] R. De Virgilio A. Maccioni, and R. Torlone. Approximate querying of rdf graphs via path alignment. Distributed and Parallel Database, pages 1-27, (2014).

[38] M. Sayyadian, H. LeKhac, A. Doan, and L. Gravano. Efficient Keyword Search Across Heterogeneous Relational Databases. In Proceedings of the 23rd International Conference on Data Engineering,ICDE 2007, April 15-20, 2007, The Marmara Hotel, Istanbul, Turkey, pages 346–355, (2007).

[39] P. Talukdar, M. Jacob, M.Mehmood, K. Crammer, Z. Ives, F. Pereira, and S. Guha. Learning to create data-integrating queries. In VLDB, (2008).

[40] S. Bergamaschi, E. Domnori, F. Guerra, M. Orsini, R.T. Lado, and Y. Velegrakis. Keymantic: semantic keyword-based searching in data integration systems. In VLDB, pages 1637-1640, (2010).

[41] V. Kantere, G. Orfanoudakis, A. Kementsietsidis, ans T. Sellis. Query relaxiation across heterogeneous data sources. In CIKM, pages 473-482, (2015).

[42] J. Li, C. Liu, J. Yu, and R. Zhou. Efficient top-k search across heterogeneous XML data sources. In International Conference on Database Systems for Advanced Applications, Springer, pages 314-329, (2008).

[43] Geist, Ingolf. Keyword Search across Distributed Heterogeneous Structured Data Sources Dissertation, Otto-von-Guericke-Universität Magdeburg, (2012).

[44] P. Talukdar, Z. Ives, and F. Pereira. Automatically incorporating New Sources in keyword search-based data integration. In SIGMOD, (2010).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[45] Z. Yan, N. Zheng, Z. Ives, P. Talukdar, and C. Yu. Actively Soliciting feedback for query answers in keyword search-based data integration. In VLDB, (2013).

[46] DBLP. http://dblp.uni-trier.de/, (last visit: 2018).

[47] N. Mamoulis, L. Yiu, H. Cheng, and W. Cheung. Efficient top-k aggregation of ranked inputs. ACM Transactions on Database Systems (TODS), 32(3), p.19, (2007).

[48] Q. Vu, B. Ooi, D. Papadias, and A.Tung. A graph method for keyword-based selection of top-k databases. In SIGMOD, pages 915-926, (2008).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Stock Market Data Analysis and Future Stock Prediction using Neural Network

Tapashi Gosswami
Department of Computer Science and Engineering
Comilla University, Comilla, Bangladesh
tapashi.cse@gmail.com

Sanjit Kumar Saha
Assistant Professor
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka, Bangladesh
sanjit@juniv.edu

Mahmudul Hasan
Assistant Professor
Department of Computer Science and Engineering
Comilla University, Comilla, Bangladesh
mhasanraju@gmail.com

***Abstract*— Share market is one of the most unpredictable and place of high interest in the world. There are no significant methods exist to predict the share price. Mainly people use three ways such as fundamental analysis, statistical analysis and machine learning to predict the share price of share market but none of these methods are proved as a consistently acceptable prediction tool. So developing a prediction tool is one of the challenging tasks as share price depends on many influential factor and features. In this paper, we propose a robust method to predict the share rate using neural network based model and compare how it differ with the actual price. For that we collect the share market data of last 6 months of 10 companies of different categories, reduce their high dimensionality using Principal Component Analysis (PCA) so that the Backpropagation neural network will be able to train faster and efficiently and make a comparative analysis between Dhaka Stock Exchange (DSE) algorithm and our method for prediction of next day share price. In order to justify the effectiveness of the system, different test data of companies stock are used to verify the system. We introduce a robust method which can reduce the data dimensionality and predict the price based on artificial neural network.**

***Index Terms*—PCA, Artificial Neural Network, Stock Market, Stock Market Prediction, DSE**

I. INTRODUCTION

Predicting anything is the most mysterious and toughest task in our world. Good prediction makes things good and bad prediction makes a huge loss. Stock market prediction is one of the toughest tasks for everyone who deals with it. Prediction with 100% accuracy is quite impossible. Good prediction means prediction with good average calculation. When someone's prediction is better at average, then he/she is a good analyst. From the beginning of world it has been our common goal to make our life easier and comfortable. The prevailing notion in society is that wealth brings comfort and luxury, there has been so much work done on ways to predict the stock markets. Various methods, techniques and ways have been proposed and used with varying results. Stock market prediction is to predict the future stock using the market statistics of past years. However, no technique or combination of techniques has been successful enough to consistently "beat the market". In my research work I have used neural network, as it is the most powerful tool to predict and analyze data. The concept of the neural network comes from the concept of our biological brain. It is very good at recognizing complex pattern and find out the unknown relation among different variables of data.

In this paper, we studied lot about the stock market statistics. We use Dhaka Stock Exchange, Bangladesh as our data source. We select 10 companies from different category and collect their last 6 months data. This data archive contain huge amount of data with multiple dimension. As a researcher we apply a statistical tool Principal Component Analysis known as PCA to reduce the data dimension. Reducing data dimension is necessary because large dataset required more time to train in neural network. After reducing data dimension we implement neural network to train the data set and neural network find the relation between different variables. After successful training we are able to test our network using existing data how well it can predict using different plotting and using different diagram. We calculate the error rate and how much data are predicted in quietly near to the original data.

II. SYSTEM ARCHITECTURE

Analysis of the huge amount of data of stock market is the main challenge for us as there are large number of organizations and company involved in stock exchange. As our goal is to predict the market price and compare those with stock market's own algorithm and testing that whether our system works better or not. The training process for the Backpropagation neural network is competitive. First of all we apply Principal Component Analysis (PCA) technique on data to reduce dimension. After reducing data dimension, we fed it in neural network for training. One neuron will "win" for each training set and it will have its weight adjusted so that it will react even more strongly to the input in the next time. As for different training set, different neurons win and their ability to

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

recognize that particular set will be increased. Now we are describing the procedure used in our stock market analyzing system.

i) Study the "Dhaka Stock Exchange" and collect the previous stock of 10 different categories organization.
ii) Store their six months data in Excel sheet.
iii) This data has multiple dimensionalities.
iv) Reduce the dimensionality using Principal Component Analysis.
v) Implement Dhaka Stock Exchange algorithm known as DSE algorithm.
vi) Train the reduced data set using neural network.
vii) Evaluate the performance of our proposed system.
viii) Compare between our algorithm and DSE Algorithm.
ix) Experimental Result Analysis

The effectiveness of the algorithm has been justified by using different organizations data. The Experiments are carried out on AMD A8-6120 2GHz PC with 8 GB RAM. The algorithm has been implemented in MATLAB 2016.

Figure 1 shows the sequence of steps of our system.

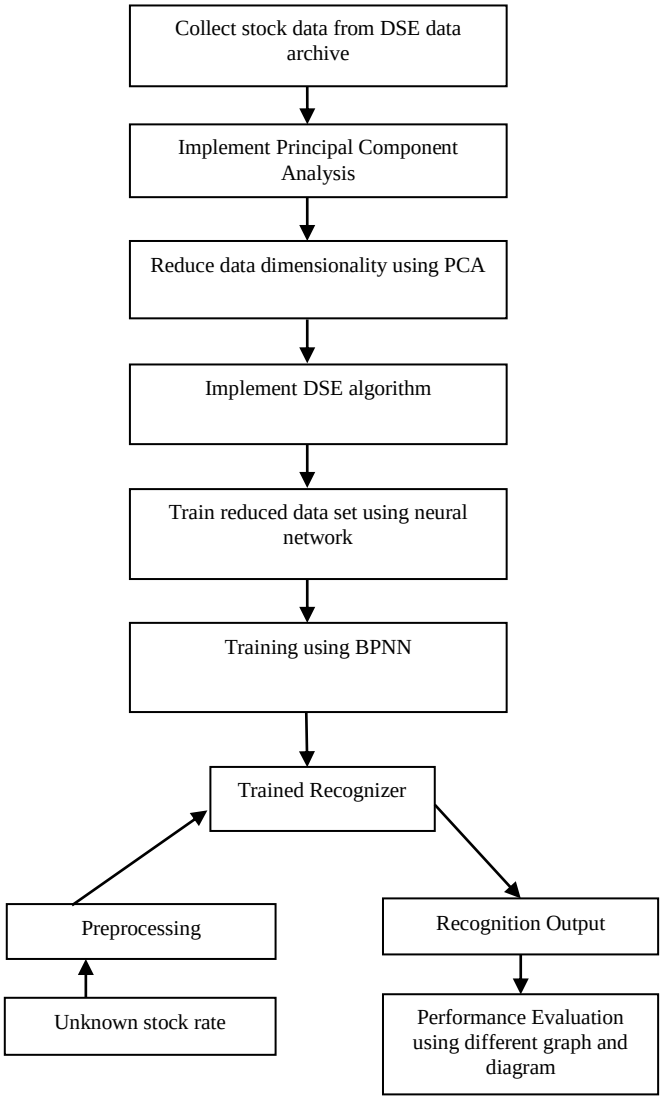


Figure 1. Overall system architecture

III. PROCESSING

A. Acquisition of Data from Stock exchange

As we study and work with stock market data, at first we need a plenty of previous data of stock market. We have studied and analyzed the "Dhaka Stock Exchange" as a data source for our research work. We select "City Bank", "ACI", "GrameenPhone", "AZIZPIPES", "BANGAS", "BEXIMCO", as our data source. After studying these companies we store their last 6 months data on Excel sheet for further processing.

A	B	C	D	E	F	G	H	I	J	K	L
#	DATE	TRADING CODE	LTP*	HIGH	LOW	OPENP*	CLOSEP*	YCP	TRADE	VALUE (mn)	VOLUME
1	04/07/2018	GP	375.5	381.9	374.3	378	375.4	376.7	1,319	55.393	146,816
2	03/07/2018	GP	375	387	375	381	376.7	381	1,652	93.922	247,529
3	02/07/2018	GP	382	390.1	380	390.1	381	388.9	1,286	77.214	202,220
4	28/06/2018	GP	389	398	384.6	398	388.9	394.6	2,076	155.466	398,540
5	27/06/2018	GP	395	401.9	393.7	400	394.6	397.8	1,376	136.372	342,968
6	26/06/2018	GP	398.3	404.9	395.2	404	397.8	403.6	1,760	200.802	503,578
7	25/06/2018	GP	405.9	411	402	408.9	403.6	408.6	2,238	214.155	526,733
8	24/06/2018	GP	409.2	415	403.5	405	408.6	403.4	3,238	213.519	522,536
9	21/06/2018	GP	399.3	405	385	385	403.4	383.2	4,074	491.023	1,235,015
10	20/06/2018	GP	382.6	385.9	379.9	381.8	383.2	379.3	1,651	143.168	373,893
11	19/06/2018	GP	379.7	384	378.8	382.6	379.3	378.8	2,121	178.529	468,335
12	18/06/2018	GP	380.6	382	374.6	374.6	378.8	374.6	2,768	285.529	756,907
13	12/06/2018	GP	374.6	381.4	373.3	381	374.6	377	2,216	146.566	389,429
14	11/06/2018	GP	376	395	376	394.1	377	390.9	2,714	166.798	435,577
15	10/06/2018	GP	390.6	398.5	388	395	390.9	395.3	1,232	55.37	141,661
16	07/06/2018	GP	394.8	407.9	392.2	406.9	395.3	405	2,004	102.36	255,916
17	06/06/2018	GP	404	408.5	403.5	407	405	405.8	1,335	102.578	252,544
18	05/06/2018	GP	405.1	410	405	409	405.8	407.6	1,359	118.691	291,729
19	04/06/2018	GP	407	412.8	401	405	407.6	402.4	1,267	92.558	227,555
20	03/06/2018	GP	400.1	415.2	400.1	415.2	402.4	414.8	1,552	89.948	220,524
21	31/05/2018	GP	416	427.9	412.7	427.9	414.8	426.3	1,735	104.091	249,254
22	30/05/2018	GP	424	428.8	424	428.4	426.3	428	966	66.619	155,972
23	29/05/2018	GP	429.8	433	426.5	432	428	431.9	798	110.136	255,330
24	28/05/2018	GP	432	439.8	431.5	433	431.9	433	1,039	122.845	282,478

Figure 2. Data archive of GrameenPhone

B. Reduction of data dimensionality

To train the network using neural network, reduction of data dimensionality is necessary for better and fast training. We have a dataset composed by a set of properties. Many of these features will measure related properties and so will be redundant. Therefore, we should remove this redundancy and describe each with less property. This is exactly what PCA aims to do. The algorithm actually constructs new set of properties based on combination of the old ones. Mathematically speaking, PCA performs a linear transformation moving the original set of features to a new space composed by principal component.

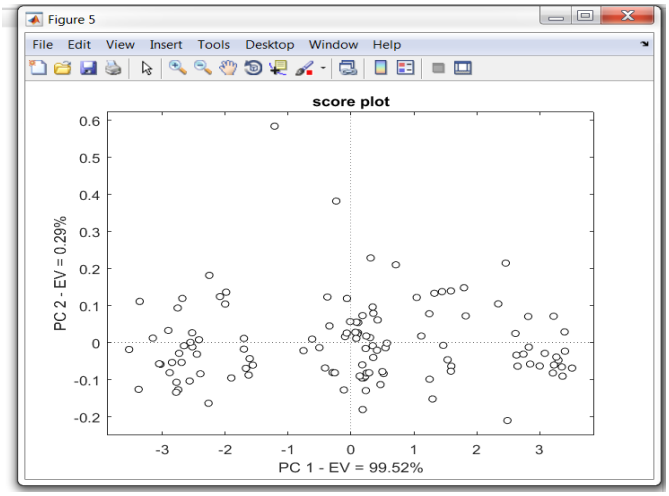


Figure 3. Score Plot of data of GrameenPhone

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

After applying PCA on our each and every dataset we got the Eigen matrix.

Editor - pca_tapashi.m

tmp_view | featues | tmp_view1

5x4 cell

	1	2	3	4	5	6
1	compone...	eigenvalue	explaine...	cumulati...		
2	PC1	3.9808	99.5208	99.5208		
3	PC2	0.011791	0.29479	99.8156		
4	PC3	0.0064981	0.16245	99.9781		
5	PC4	0.000876...	0.021912	100		
6						
7						
8						

Figure 4. Eigen Matrix

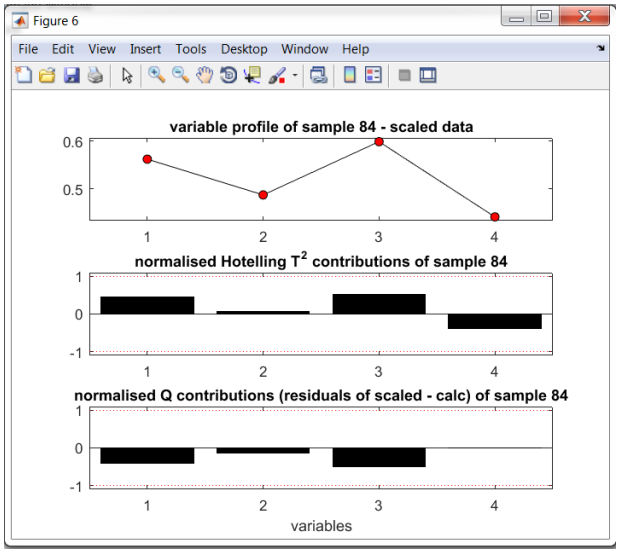


Figure 5. Normalized values for properties

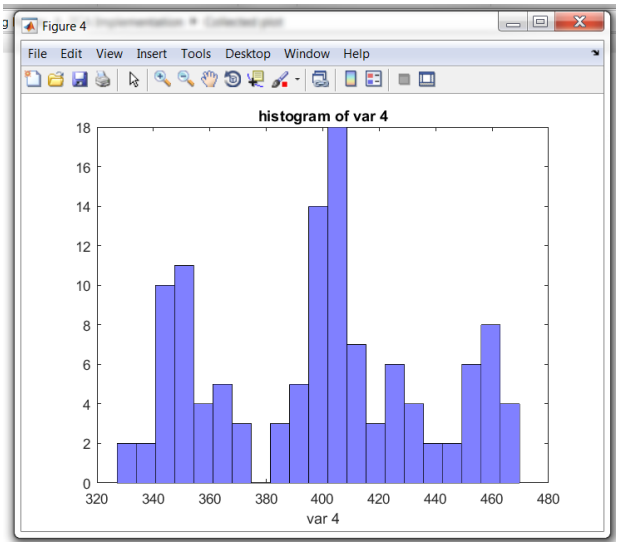


Figure 6. Sample histogram for variable 4

For the data of our each selected company we apply PCA on them and get the most influential properties. After successful implementation of Principal Components Analysis we write the reduced data on Microsoft Excel sheet. Now our data are prepared as an input for neural network.

Figure 7 shows the data archive before Applying PCA on GrameenPhone company data:

A	B	C	D	E	F	G	H	I	J	K	L
#	DATE	TRADING CODE	LTP*	HIGH	LOW	OPENP*	CLOSEP*	YCP	TRADE	VALUE (mn)	VOLUME
1	04/07/2018	GP	375.5	381.9	374.3	378	375.4	376.7	1,319	55.393	146,816
2	03/07/2018	GP	375	387	375	381	376.7	381	1,652	93.922	247,529
3	02/07/2018	GP	382	390.1	380	390.1	381	388.9	1,286	77.214	202,220
4	28/06/2018	GP	389	398	384.6	398	388.9	394.6	2,076	155.466	398,540
5	27/06/2018	GP	395	401.9	393.7	400	394.6	397.8	1,376	136.372	342,968
6	26/06/2018	GP	398.3	404.9	395.2	404	397.8	403.6	1,760	200.802	503,578
7	25/06/2018	GP	405.9	411	402	408.9	403.6	408.6	2,238	214.155	526,733
8	24/06/2018	GP	409.2	415	403.5	405	408.6	403.4	3,238	213.519	522,536
9	21/06/2018	GP	399.3	405	385	385	403.4	383.2	4,074	491.023	1,235,015
10	20/06/2018	GP	382.6	385.9	379.9	381.8	383.2	379.3	1,651	143.168	373,893
11	19/06/2018	GP	379.7	384	378.8	382.6	379.3	378.8	2,121	178.529	468,335
12	18/06/2018	GP	380.6	382	374.6	374.6	378.8	374.6	2,768	285.529	756,907
13	12/06/2018	GP	374.6	381.4	373.3	381	374.6	377	2,216	146.566	389,429
14	11/06/2018	GP	376	395	376	394.1	377	390.9	2,714	166.798	435,577
15	10/06/2018	GP	390.6	398.5	388	395	390.9	395.3	1,232	55.37	141,661
16	07/06/2018	GP	394.8	407.9	392.2	406.9	395.3	405	2,004	102.36	255,916
17	06/06/2018	GP	404	408.5	403.5	407	405	405.8	1,335	102.578	252,544
18	05/06/2018	GP	405.1	410	405	409	405.8	407.6	1,359	118.691	291,729
19	04/06/2018	GP	407	412.8	401	405	407.6	402.4	1,267	92.558	227,555
20	03/06/2018	GP	400.1	415.2	400.1	415.2	402.4	414.8	1,552	93.948	220,524
21	31/05/2018	GP	416	427.9	412.7	427.9	414.8	426.3	1,735	104.091	249,254
22	30/05/2018	GP	424	428.8	424	428.4	426.3	428	966	66.619	155,972
23	29/05/2018	GP	429.8	433	426.5	432	428	431.9	798	110.136	255,330
24	28/05/2018	GP	432	439.8	431.5	433	431.9	433	1,039	122.845	282,478

Figure 7. Data archive of GrameenPhone

After applying PCA:

A	B	C	D	E	F	G	H
#	DATE	TRADING CODE	LTP*	HIGH	LOW	OPENP*	CLOSEP*
1	04/07/2018	GP	375.5	381.9	374.3	378	375.4
2	03/07/2018	GP	375	387	375	381	376.7
3	02/07/2018	GP	382	390.1	380	390.1	381
4	28/06/2018	GP	389	398	384.6	398	388.9
5	27/06/2018	GP	395	401.9	393.7	400	394.6
6	26/06/2018	GP	398.3	404.9	395.2	404	397.8
7	25/06/2018	GP	405.9	411	402	408.9	403.6
8	24/06/2018	GP	409.2	415	403.5	405	408.6
9	21/06/2018	GP	399.3	405	385	385	403.4
10	20/06/2018	GP	382.6	385.9	379.9	381.8	383.2
11	19/06/2018	GP	379.7	384	378.8	382.6	379.3
12	18/06/2018	GP	380.6	382	374.6	374.6	378.8
13	12/06/2018	GP	374.6	381.4	373.3	381	374.6
14	11/06/2018	GP	376	395	376	394.1	377
15	10/06/2018	GP	390.6	398.5	388	395	390.9
16	07/06/2018	GP	394.8	407.9	392.2	406.9	395.3
17	06/06/2018	GP	404	408.5	403.5	407	405
18	05/06/2018	GP	405.1	410	405	409	405.8
19	04/06/2018	GP	407	412.8	401	405	407.6
20	03/06/2018	GP	400.1	415.2	400.1	415.2	402.4
21	31/05/2018	GP	416	427.9	412.7	427.9	414.8

Figure 8. After applying PCA on data of GrameenPhone

IV. PERFORMANCE EVALUATION AND EXPERIMENT RESULT ANALYSIS

A. Implement Dhaka Stock Exchange (DSE) Algorithm

At the beginning we discuss about the different types of prediction, fundamental analysis is one of them. As we select Dhaka Stock Exchange as our data source, we find they have own algorithm for stock rate prediction. To compare it with our system at first we implement the DSE algorithm in MATLAB and write the answer and predicted value in another file. The short description of DSE Algorithm as follows.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

LTP = Last Traded Price
CLOSEP = Closing Price
YCP = Yesterday's Closing Price
OAP = Open Adjusted Price

Index calculation algorithm (according to IOSCO Index Methodology):

$$\text{Current Index} = \frac{\text{Yesterday's Closing Index X Current M.Cap}}{\text{Opening M.Cap}}$$

$$\text{Closing Index} = \frac{\text{Yesterday's Closing Index X Closing M.Cap}}{\text{Opening M.Cap}}$$

Current M.Cap = ∑ (LTP × Total no. of indexed shares)
Closing M.Cap = ∑ (CP × Total no. of indexed shares)

Abbreviations and Acronyms

M.Cap - Market Capitalization
DSE - Dhaka Stock Exchange
IOSCO - International Organization of Securities Exchange Commissions (IOSCO)
LTP - Last Traded Price
CP - Closing Price

Here we give a sample example of GrameenPhone after implementing DSE algorithm for next day prediction:

#	LTP*	HIGH	LOW	OPENP*	CLOSEP*	YCP	CurrentIndex	ClosingIndex
1	375.5	381.9	374.3	378	375.4	376.7	291.9244	291.979
2	375	387	375	381	376.7	381	292.7424	292.7977
3	382	390.1	380	390.1	381	388.9	293.567	293.6138
4	389	398	384.6	398	388.9	394.6	294.3634	294.415
5	395	401.9	393.7	400	394.6	397.8	295.1301	295.1833
6	398.3	404.9	395.2	404	397.8	403.6	295.8263	295.8832
7	405.9	411	402	408.9	403.6	408.6	296.4874	296.5509
8	409.2	415	403.5	405	408.6	403.4	297.0675	297.16
9	399.3	405	385	385	403.4	383.2	297.6149	297.7164
10	382.6	385.9	379.9	381.8	383.2	379.3	298.3445	298.3271
11	379.7	384	378.8	382.6	379.3	378.8	299.5404	299.5172
12	380.6	382	374.6	374.6	378.8	374.6	300.7581	300.7391
13	374.6	381.4	373.3	381	374.6	377	301.9296	301.9454
14	376	395	376	394.1	377	390.9	303.2526	303.2687
15	390.6	398.5	388	395	390.9	395.3	304.5443	304.549
16	394.8	407.9	392.2	406.9	395.3	405	305.6496	305.6531
17	404	408.5	403.5	407	405	405.8	306.7509	306.751
18	405.1	410	405	409	405.8	407.6	307.7883	307.7813
19	407	412.8	401	405	407.6	402.4	308.8234	308.8106
20	400.1	415.2	400.1	415.2	402.4	414.8	309.845	309.8282
21	416	427.9	412.7	427.9	414.8	426.3	310.9064	310.8747
22	424	428.8	424	428.4	426.3	428	311.8753	311.852
23	429.8	433	426.5	432	428	431.9	312.7805	312.7464
24	432	439.8	431.5	433	431.9	433	313.6602	313.6394

Figure 9. Result after DSE algorithm Implementation for GrameenPhone

B. *Training*

When we apply principal component analysis in our data set we got the most influential data properties. These features are used as an input vector for our neural network. For our neural network we use closing price list of last 6 months of specific organization as a target vector. The closing price is the final price at which a security is traded on a given trading day. The closing price represents the most up-to-date valuation of a security until trading commences again on the next trading day.

Hence, in our system the number of neurons in input layer is 124*5, neurons in input layer for neural network and number of neurons in hidden layer is 20, and finally the neurons in output layer is 124. The number of neurons in hidden layer can vary from 25% to 50% of its input neurons.
In our Training:

- Number of neuron in input layer : 124*5
- Number of neuron in hidden layer : 20
- Number of neuron in output layer : 124
- Learning rate (α) : 0.001 and threshold (θ) : 0.9
- Epoch : 2000

After specification of these experimental parameters we move into the implementation of Backpropagation neural network (BPNN). At first, initial weights are randomly generated for the network. In feed forward step, input patterns are propagated through the network one by one and actual outputs are calculated. Comparing between actual and target outputs, according to BPNN algorithm, magnitude of error is determined and weight updates take place through back propagation with a view to minimize the error subsequently. When all patterns in pattern set are fed into the network and weights are updated as stated earlier, this constitutes one epoch as per the definition of literature.

Updated Weight = weight (old) + learning rate * output error * output (neurons i) * output (neurons i+1) * (1 - Output (neurons i+1)).

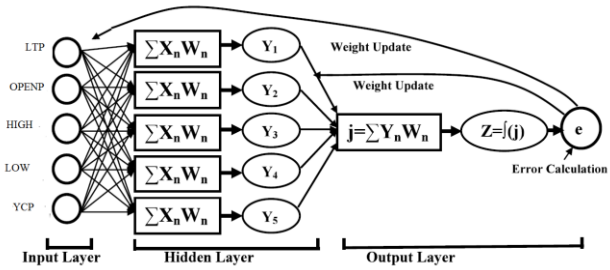


Figure 10. Network architecture

Hidden layer neuron generate the final output which is the compared with the real output and calculate an error signal *e*. Unless estimated reaches satisfactory level measured based on error threshold as specified before, epochs are continued. In this work, 10 different companies are trained individually and their results are analyzed.

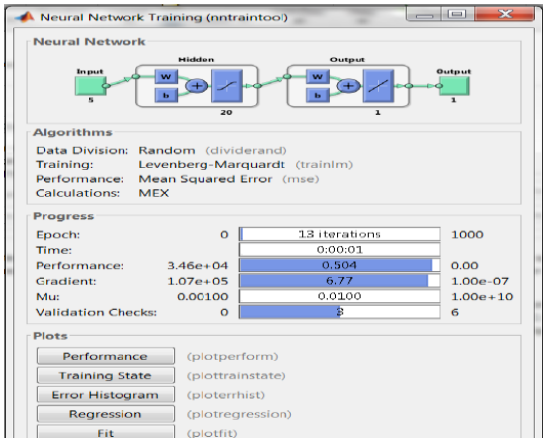


Figure 11. Training the network

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

After 1000 iteration the training is stopped and error rate is in minimum convergence level.

At 201 epochs best training performance is found.

C. *Testing*

The performance of a network is often measured based on how well the system predicts market direction. Ideally, the system should predict market direction better than current methods with less error. Some neural networks have been trained to test. If a neural network can outperform the market consistently or predict its direction with reasonable accuracy, the validity of network is questionable. Other neural networks were developed to outperform current statistical and regression techniques. Most of the neural networks used for predicting stock prices. In order to justify the performance of the neural network, various experiments are carried out. All experiments are performed with for training the system for recognition. There is no overlap between the training and test data sets. After successful completion of the training different company's data, neural network is used to recognize unknown data both as a whole and separately.

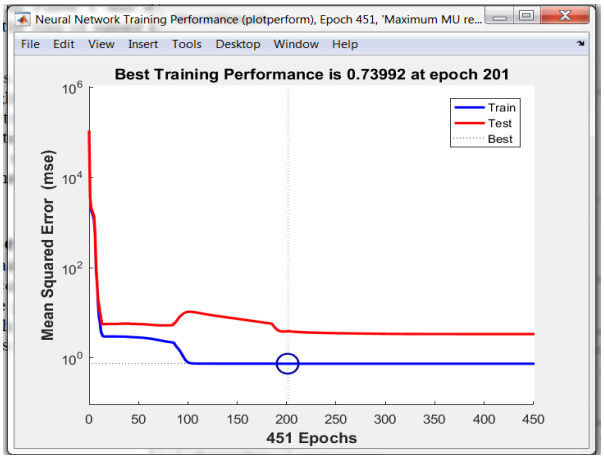


Figure 12. Best Training Performance

We have carried out the experiment for 10 different category companies and observed their last 6 months data, apply PCA on them to reduce data dimension and finally train them with neural network and predict their data depending on their previous day rate. After that we compare their performance using mean square error.

The data set is partitioned into two parts. The first one is used for training the system and the second one for test the system in order to evaluate the performance. For each organization, features are computed, reduced and stored for training the network. Three layers Backpropagation neural network (BPNN), i.e. one input layer, one hidden layer and one output layer are taken. If number of neurons in the hidden layer is increased, then a problem of allocation of required memory is occurred. Also, if the value of error tolerance is high, desired results are not obtained, so changing the value of error tolerance in a minimum value, high accuracy rate is obtained. Also the network takes more number of epochs to learn when the error tolerance value is less rather than in the case of high value of error tolerance in which network learns in less number of cycles and so the learning is not very fine. The result also varies for each organization. Various experiments are carried out to justify the performance of the system. All experiments are performed with set of training data and completely different data set for test. There is no overlap between the training and test data sets. The neural network is trained using the default learning parameter settings (learning rate 0.001, threshold 0.9) for 1000 epochs.

Table I Predicted Price of GrameenPhone

Date	LTP	Open Price	Close Price	Predicted Close Price	Error Rate (%)
04/07/2018	375.5	378	375.4	377	0.42621
05/07/2018	382	378	383.9	379.7	1.09403
08/07/2018	390	387	387.8	385.5	0.59309
09/07/2018	388	389	387.9	387.2	0.18046
10/07/2018	380	390.6	379.5	382.4	0.76416
11/07/2018	382	380	380.7	379.1	0.42028
12/07/2018	388.6	385	388.5	385.7	0.72072
15/07/2018	381	390	383.1	386.4	0.86139
16/07/2018	386.3	390	388	389.2	0.30928
17/07/2018	399	392	397.7	392	1.43324

Our prediction for GrameenPhone for 10 days in month of July shows how it closely related to our predicted price and its actual price. GrameenPhone is an 'A' category company. Following plot shows how target data fits with train data.

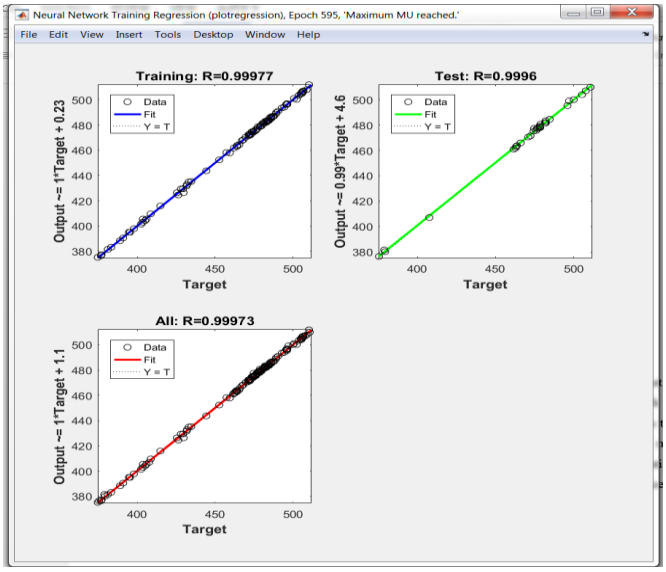


Figure 13. Snapshot of how target data fits to with the train data

Error Histogram for GrameenPhone Company is shown below:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

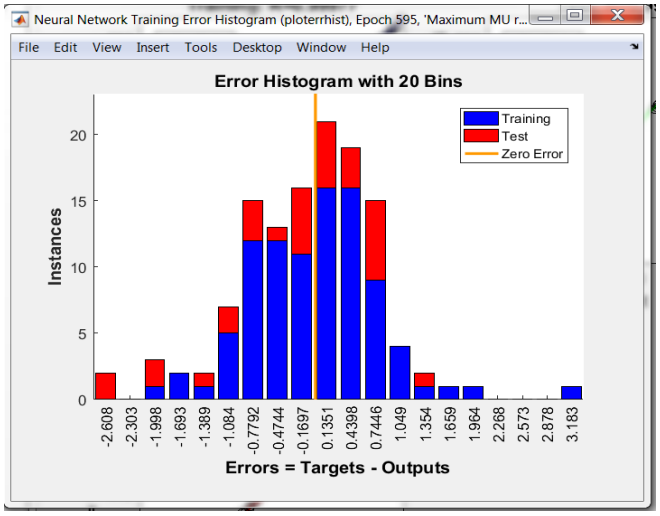


Figure 14. Error Histogram for GrameenPhone with 20 bins

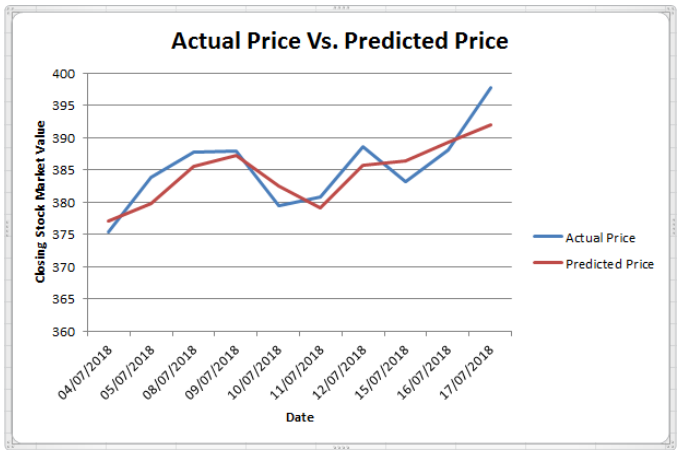


Figure 15. Predicted price of GrameenPhone for 10 days

The experimental results show that this robust method is effective and efficient in forecasting stock prices compared with DSE algorithm for stock market prediction.

Now we have considered one 'B' category company. For our research work we Choose Company named 'Azizpipes'. As like GrameenPhone we apply PCA on it for data dimension reduction, than apply DSE algorithm on it. After that we trained the reduced data using Backpropagation neural network. After successful training we test it using some sample data and compared it with its actual price.

Table II Predicted Price of AzizPipes

Date	LTP	Open Price	Close Price	Predicted Close Price	Error Rate (%)
01/02/2018	148.6	155	151	152	0.66
04/02/2018	143.5	151.9	142.3	145.5	2.25
05/02/2018	141	145	140.9	139	1.35
06/02/2018	143	144	142.4	144	1.12
07/02/2018	143.1	142.2	144.7	145	0.21
08/02/2018	143.1	145	143	140	2.10

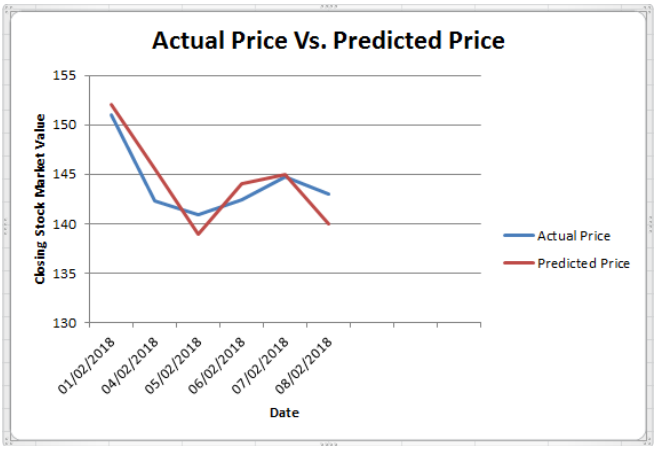


Figure 16. Predicted price of AzizPipes for 6 days

Here, we observed that for some sample our predicted price is so close to actual price.

V. CONCLUSION

The main purpose of working on stock market prediction is to increase the investor in stock market and make the significant profit by predicting the market rate. In our proposed system, we developed a model to predict the stock rate of a specific company by training their previous data in neural network. To train our system faster, we at first reduce the data dimension using PCA. After successful reduction of the data dimension we got the most influential features of data. After training, we test our system how successfully it predict the stock rate. We observed that if we use more data to train the network then the performance will be increased significantly. For our system we used Backpropagation neural network which is one of the best neural network. It reduces an error between the actual output and desired output in a gradient descent manner. Performance is not always satisfactory because it will be quite difficult to predict with 100% accuracy.

REFERENCES

[1] Ahmed, M. U.; Begum, S.; Funk, P.; Xiong, N.; and Schéele, B. von. 2008. A Three Phase Computer Assisted Biofeedback Training System Using Case-Based Reasoning, In 9th European Conference on Case-based Reasoning workshop proceedings. Trier, Germany.

[2] Ahmed, M. U.; Begum, S.; Funk, P.; Xiong, N.; and Schéele, B. von. 2008a. Case-based Reasoning for Diagnosis of Stress using EnhancedCosine and Fuzzy Similarity, Transactions on Case-Based Reasoning on Multimedia Data, vol 1, nr 1, IBaI Publishing, ISSN: 1864-9734

[3] Begum, S.; Ahmed, M. U.; Funk, P.; Xiong, N.; and Schéele, B. von. 2008. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching, In Computational Intelligence (CI), in press, Blackwell Bichindaritz, I. 2006. Case-based

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Reasoning in the Health Sciences. In Artificial Intelligence in Medicine 36(2), 121-125

[4] CaseBook (McSherry 2007) [Purpose: Diagnosis, Classification] is a hypothetico-deductive CBR system for classification and diagnosis that applies hypotheticodeductive reasoning (HDR) in conversational CBR systems.

[5] ExpressionCBR (De Paz et al. 2008) [Purpose: Diagnosis, Classification]

[6] Fungi-PAD (Perner et al. 2006, Perner and Bühring 2004) [Purpose: Classification, Knowledge acquisition/management]

[7] FrakaS (Cordier et al. 2007) [Purpose: Diagnosis, Knowledge acquisition/ management]

[8] GerAmi (Corchado, Bajo, and Abraham 2008) [Purpose: Planning, Knowledge acquisition/ management] 'Geriatric Ambient Intelligence'

[9] The KASIMIR project (D'Aquin, Lieber, and Napoli 2006) [Purpose: Diagnosis, Classification, Knowledge acquisition/ management]

[10] HEp2-PAD (Plata et al. 2008; Perner 2006a; Perner management

[11] Case-Based Reasoning System to Attention-Deficit Hyperactivity Disorder. In CBR research and development: 6th International Conference on CBR, 122-136. ICCBR'05.

[12] Cordier, A.; Fuchs, B; Lieber, J.; & Mille, A. 2007. On-Line Domain Knowledge Management for Case-Based Medical Recommendation. In Workshop on CBR in the Health Sciences, pp. 285-294. ICCBR'07

[13] Díaz, F.; Fdez-Riverola, F.; and Corchado, J.M. 2006. GENE-CBR: a Case-Based Reasoning Tool for Cancer Diagnosis using Microarray Datasets. In Computational Intelligence. Vol/Iss 22/3-4, pp. 254-268.

[14] Jolliffe, I. T. (1986). Principal Component Analysis. Springer, New York.

[15] Tapashi Gosswami and Sanjit Kumar Saha, "Handwritten Bangla Numeral Recognition Exerting Neural Network Approach", International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 4, April 2017, pp. 193-196.

AUTHORS PROFILE

Tapashi Gosswami has completed her Bachelor of Science in Computer Science and Engineering from Comilla University, Bangladesh. Currently, Ms. Gosswami is studying her Master of Science in the department of Computer Science and Engineering, Comilla University, Bangladesh. Gosswami has published journal paper in the International journal. Her research interest includes artificial neural network, machine learning, image processing.

Sanjit Kumar Saha has obtained his both Bachelor of Science and Master of Science degree in computer science and engineering from Jahangirnagar University, Bangladesh in 2007 and 2009 respectively. He is currently working as an assistant professor of department of computer science and engineering at Jahangirnagar University, Bangladesh. He has 8+ years of teaching experience to both undergraduate and graduate students. Saha has published journal and conference papers in the International and National journal and conferences. His current interest includes deep learning, artificial neural network, fuzzy logic and systems.

Mahmudul Hasan who obtained an M.Sc.(Thesis) in Computer Science and Engineering from University of Rajshahi, Bangladesh. He is currently working as a faculty (Assistant Professor) in the Department of Computer Science and Engineering at Comilla University, Comilla, Bangladesh. His teaching experience includes different graduate (M.Sc.) and under graduate courses along with many thesis and project supervision. He is a member of various National/International associations. He is the reviewer of the different national and international conferences and journals.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Iris Segmentation by Using Circular Distribution of Angles in Smartphones Environments

Rana Jassim Mohammed[a],Taha Mohammad Al Zaidy[b] and Naji Mutar Sahib[c]

[a]Department of Computer Sciences, College of Science, University of Diyala, Diyala, Iraq
rana.jassim0@gmail.com

[b]*Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq*
taha_alzaidy@yahoo.com

[c]*Department of Computer Sciences, College of Science, University of Diyala, Diyala, Iraq*
naji_e2006@yahoo.com

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ABSTRACT

The widespread use of smartphones with internet connectivity has resulted in the storage and transmission of sensitive data. This has heightened the need to perform reliable user authentication on smartphones in order to prevent an adversary from accessing such data. Biometrics, the science of recognizing individuals based on their biological and behavioural traits. In order to perform reliable verification in smartphones, we briefly discuss the suitability of using the iris texture for biometric recognition in smartphones. One of the critical components of an iris recognition system is the segmentation module which separates the iris from other ocular attributes. In this paper, we propose a new and robust iris segmentation method based on circular distribution of angles to localize the iris boundary, which applied on an eye image passed through the pre-processing operation . The experiment results are carried out on the MICHE-I (Mobile Iris Challenge Evaluation) dataset Samsung Galaxy S4 (SG4) iris image database. The evaluation of the obtained results shows that the developed system can successfully localize iris on the tested images under difficult environments compared to previous techniques and we have achieved 85% Av accuracy rate with (SG4).

Keywords: Biometric, Iris Segmentation , Circular Distribution of Angles, MICHE-I Dataset.

Introduction

The mobile phones that are able to be used as small computers and connect to the internet are referred to as “smartphones”. Although, about 20 years ago the first smartphone - IBM Simon – was presented, yet the world has witnessed a massive evolution of smartphones occurred when the first iPhone was introduced in 2007.

Smartphones that are currently carried by humans are not only computers, databases, phones, cameras, locators, infinite jukeboxes and have the information in the world at humans’ fingertips, but also personal companion which is a part of their everyday life. The role of smartphones as guardians, helpers, and companions is speculated grow much bigger [1].

The number of mobile phone users worldwide is expected to pass the five billion mark by 2019, and by 2018, the number of tablet computer users is projected to reach 1.43 billion. This proliferation of smartphones and tablets raises concerns about the security and privacy of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

data stored on mobile devices if they are lost, stolen, or hacked. An attacker with physical access to a mobile device can potentially steal a user's banking information, read his/her emails, look at his/her private photos and perform other criminal actions. The scale of the problem is vast; according to ConsumerReports.org, 2.1 million mobile phones were stolen and 3.1 million phones were lost in 2013 in the United States alone [2].

For the purpose of balancing the user convenience with the necessary security by these smartphones, a mechanism is used to authenticate the users regularly with something with the users are or what users doing. This attribute should be unique for each user and not to be easy to mimic or stolen. It's possible to use biometrics for this purpose [1].

The science of persons' recognition based on their behavioural or physical attributes is referred to biometrics. Some behavioural characteristics that are used to identify users are keystroke dynamics, signatures and voice patterns, while physical characteristics include retinal patterns, palm prints, hand geometry, fingerprints and iris patterns [3]. Among the different types of physical characteristics that are available, the iris is the most common and highly adopted biometric attributes because of its universality, accuracy, and persistence. Recognition of iris is the method of identifying users based on their iris pattern [4].

In this paper , we proposes a new and robust iris segmentation method for unconstrained visible spectrum iris recognition specifically tailored for the smartphone based iris recognition applications. The iris segmentation scheme presented in this work relies on circular distribution of angles as a basis to localize the iris boundary in difficult environments. The proposed method is evaluated on the publicly available smartphone iris dataset from BIPLab1.

The rest of this paper is organized as follows: in section 2 main works related to this research are briefly recalled , in section 3 the proposed method is described with regard to iris segmentation . Section 4 describes the MICHE-I dataset and experimental results. Finally, Section 5 include conclusion and future work.

Related Work

Several methods have been proposed for iris segmentation in smartphones over the last few years:

Kiran B.Raja et al.[5],2014: Propose a new segmentation scheme and adapt it to smartphone-based visible iris images for approximating the radius of the iris to achieve robust segmentation. The proposed technique has shown the improved segmentation accuracy by up

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

to 85% with standard OSIRIS v4.1. To evaluate the proposed segmentation scheme and feature extraction scheme, we employ a publicly available database and also compose a new iris image database. The newly composed iris image database (VSSIRIS) is acquired using two different smartphones – iPhone 5S and Nokia Lumia 1020 under mixed illumination with unconstrained conditions in the visible spectrum. The biometric performance is benchmarked based on the equal error rate (EER) obtained from various state-of-art schemes and the proposed feature extraction scheme. An impressive EER of 1.62% is obtained on our VSSIRIS database and an average gain of around 2% in EER is obtained on the public database as compared to the well-known state-of-art schemes.

Andrea F. Abate et al. [6], 2014: Proposed technique is based on watershed transform for iris detection in unclear images captured with the use of mobile devices. This method takes the advantage of information related to limbus for segmenting the and merging its scores with the iris' once in order to have a more accurate recognition phase, BIRD has been examined on iris images contained in the MICHE dataset, these images have been captured using three different mobile devices both in outdoor and in indoor settings. A comparison has been held for the results and those provided by two states of the art techniques, namely ISIS and NICE-I, it showed that iris detection\ recognition represents a challenging task, but underlining the high potentials of the proposed technique.

S. Memar Zadeh and A. Harimi [7],2016: A new iris localization method is presented for mobile devices. The proposed system uses both the intensity and saturation thresholding on the captured eye images to determine the iris boundary and sclera area, respectively. The estimated iris boundary pixels placed outside the sclera are removed. The remaining pixels are mainly the iris boundary inside the sclera. Then a circular Hough transform is applied to such iris boundary pixels in order to localize the iris. The experiments are carried out on 60 iris images taken by an HTC mobile device from 10 different persons with both the left and right eye images available per person. We also evaluate the proposed algorithm on the MICHE datasets for iphone5, Samsung Galaxy S4, and Samsung Galaxy Tab2. The evaluation results obtained show that the developed system can successfully localize iris on the tested images.

Narsi Reddy et al. [8],2016: Proposes an iris segmentation algorithm for the visible spectrum. It is based on combining Daugman's integro- differential algorithm and K-means clustering. The efficiency of this proposed approach has been proved by the experimental investigations on the publicly available VISOB dataset. Experimental results reveal that iris segmentation

increases 4 folds and 3.5 folds compared to Daugman's and Masek's methods respectively. The proposed method also executes 8 times faster than Masek's methods and 5 times faster than Daugman's.

A. Radman et al. [9], 2017: A new iris segmentation method is developed and tested on UBIRIS.v2 and MICHE iris databases that reflect the challenges in recognition by unconstrained images. This method accurately localizes the iris by a model designed on the basis of the Histograms of Oriented Gradients (HOG) descriptor and Support Vector Machine (SVM), namely HOG-SVM. Based on this localization, iris texture is automatically extracted by means of cellular automata which evolved via the Grow Cut technique. Pre- and post-processing operations are also introduced to ensure higher segmentation accuracy. Extensive experimental results illustrate the effectiveness of the proposed method on unconstrained iris images.

The Proposed Method

As mentioned previously accurate iris localization plays a significant role in improving the performance of iris recognition systems. The goal of iris segmentation is the precise detection of outer boundaries of the iris region. In this paper, the proposed iris segmentation method is based on the circular distribution of angles. The diagram of the proposed method contains three stages: Eye image acquisition, Eye image pre-processing and Iris localization as shown in (figure 1).

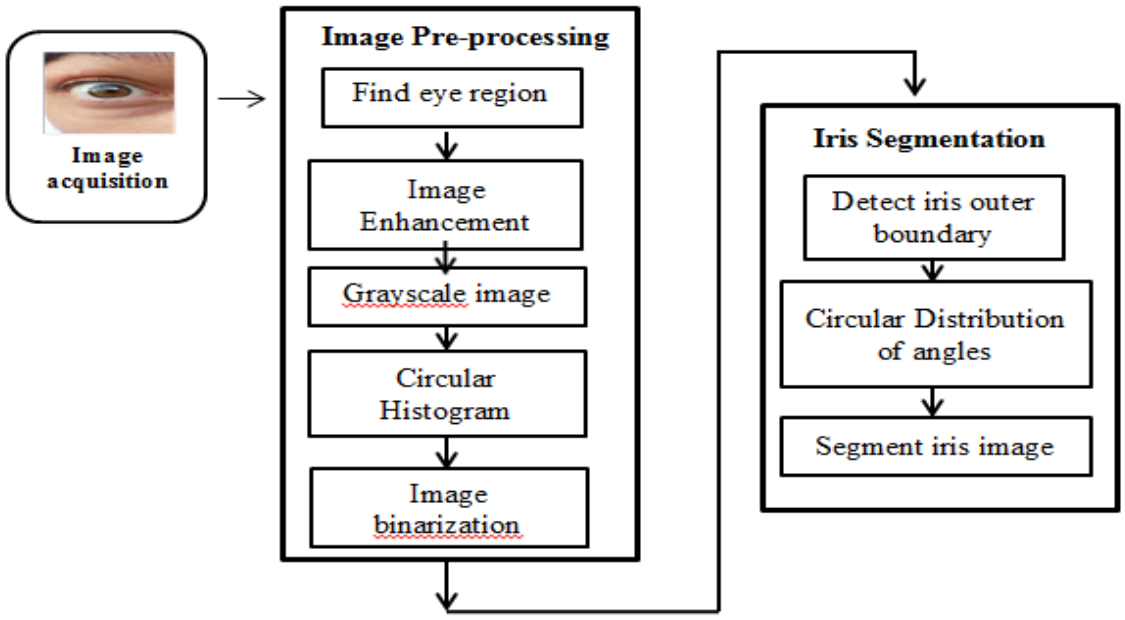


Figure 1: Block diagram of the proposed iris segmentation

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Image acquisition stage

Eye Image Acquisition is the first stage in the system, this stage depends on how much clarity and purity of the captured image are provided. In this work , a specific class of iris images dataset is provided from the data set (MICHE)-I Mobile Iris Challenge Evaluation , MICHE database contains photos captured indoor and outdoor with three different mobile devices: Samsung Galaxy S4 (hereinafter GS4), iPhone 5 (hereinafter iP5) and (Samsung Galaxy Tab 2), Both the front and the rear cameras of this device is used. As we performed iris segmentation, among the three devices, we selected the highest resolution cameras: GS4.

Image Pre-processing Stage

One of the major issues in the iris segmentation and recognition system is the preprocessing. The pre-processing stage is necessary to prepare the iris image for further processing. The main goal of this stage is to localize the iris region from the eye image which can result in the accurate iris localization and there by lead to excellent iris recognition system performance.

In this paper , the image preprocessing consists of five steps: Find eye region, enhance eye image, convert the color image to the grayscale image, apply circler histogram to find the initial centre of the iris image**,** Finally, apply the global threshold to convert Grayscale image to binary image. Figure (2) illustrates the preprocessing stage.

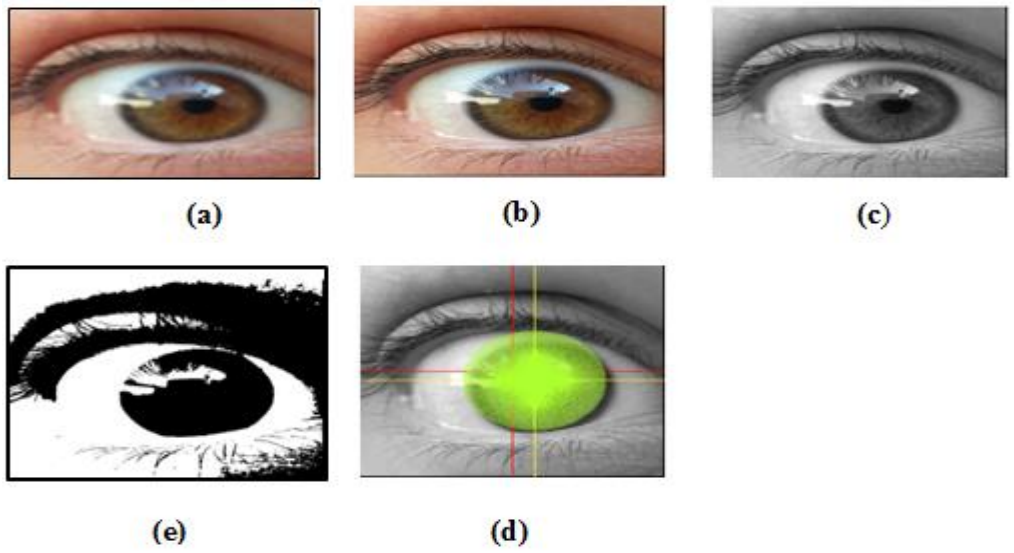


Figure 2: Pre-processing stage :(a) Eye region ,(b) Image contrast ,(c) Gray scale image ,(d) Apply circular histogram ,(e) Apply global threshold

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The first task of the pre-processing stage is to locate the eye region. To achieve this, the haar cascade library is used as eye object detector one of the tools of Open CV is implemented using C# language as shown in figure 3.

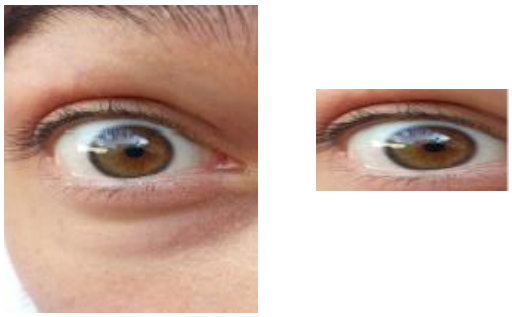
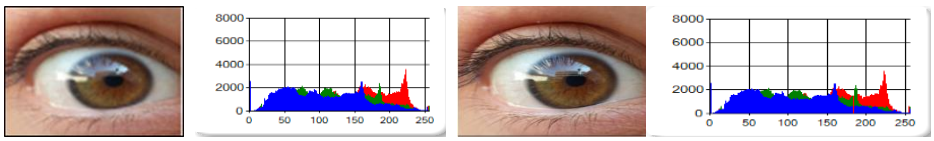


Figure 3: Example of find eye region results using Galaxy S4
,(a) Original image ,(b) Find eye region

The next step included performing image contrast stretching using parameter value between the range [0.0 - 3.0] , 0.9 is the best coefficient that can be used for most images according to experience and make eye image region more distinguishable as shown in Figure (4). The low contrast between pupil and iris is a noise factor that causes degradation in segmentation accuracy. To overcome this issue, enhancing the contrast is required before proceeding further to the iris segmentation process .

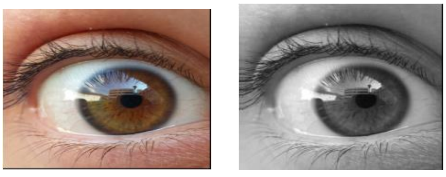


(a) **(b)** **(c)** **(d)**

Figure (4): Examples of Contrast Stretching ,(a) Original image , (b) Histogram of original image ,(c) Image contrast, (d) Histogram of image contrast .

Converting the colour image to grayscale image by using a sophisticated version of the average method is the luminosity method. It also averages the values as shown in figure 5. The formula of luminosity is:

$$0.21\ R + 0.72\ G + 0.07\ B. \qquad (1)$$



(a) **(b)**

Figure 5: Examples of image conversation ,
(a) Image after contrast stretching ,(b) Grayscale image.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The fourth stage To find the initial center of iris image using the circular histogram. The histogram of a an image computed in the range (0 – 255), in this paper circler histogram is computed in a range (0 -30) colors depending on the center of the original image based on equations (2) and (3).

with a default radius of 120 and angle from(0 to 360) we use the circle equation to draw a circle and draw the histogram of this space which we take the lowest occurrences number not equal to zero in arrange (0-30) color to find the initial center of the iris regionas shown in figure 6.

xc = Height of image / 2 (2)

yc= Width of image / 2 (3)

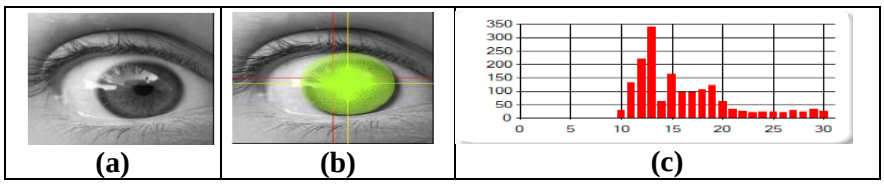


Figure 6:An Example of Iris Eye Circular Histogram (a) Grayscale image, (b) Circular histogram Initial (xc ,yc)=(256,256), New (xc ,yc) =(310,286) , (c) histogram in arrange (0-30)

The last stage in the a pre-processing , apply global threshold creates binary images from gray-level ones by turning all pixels below some threshold to zero and all pixels about that threshold to one. In our proposed system, the proper value of the threshold (***T***) is estimated to be (127)as shown in figure(7).

Figure 7: Examples of Image conversation ,(a) Gray scale image , (b) global threshold with best threshold T=126

Iris Segmentation stage

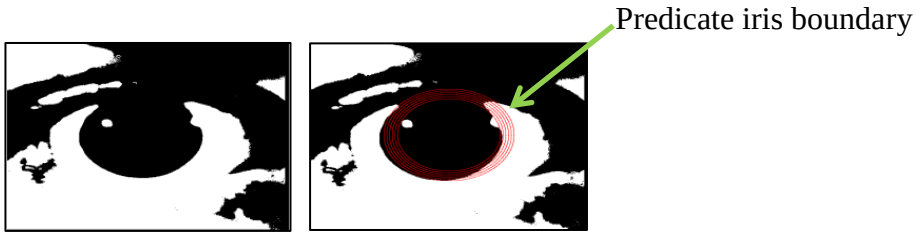
The most important stage in iris suggested system is the iris segmentation due to all the sub stages based on the accuracy of this stage. This stage is applied for localizing and extracting the iris area from the image. The iris has a ring shape, could be roughly modeled as an annulus. Iris segmentation should be accomplished with high accuracy because it significantly affects the success of the system. To make the process simple and accurate, there

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

are two steps required to segment iris region. Firstly, predicate iris outer boundary, secondly, circular distribution of angles is applied to the segmented iris image.

Step 1: Predicate Iris Outer Boundary

The 1st step in segmentation stage ,to predicate the iris boundary after getting: (i) the approximate location of the iris center (xc, yc) which is computed from Circular Histogram as the initial center and initial radius (120 to 0) with angle from (0 to 360) , we use the circle equation to draw a circle , this step is repeated till reaching the case that a significant ratio of circle pixels are black is equal to 100%, new radius is calculated. the objective of predicate iris boundary algorithm is to asses more accurately the circle parameters that fit the collected of iris's segment as shown in figure 8.



(a) (b)

Figure 8:An example of Predicate Iris Boundary , (a) Binary image, (b) Predicate Iris Outer Boundary

Step2 : Circular Distribution of Angles

Since the iris was identified as having the largest black area (based on the number of pixels) in the binary image. So, the proposed method makes a scan to allocate the largest black segment in the binary image.

After drawing the red circles and predicate iris outer boundary in previous step , In this step , firstly removing the existing specular spot reflections within the specified area ,secondly the directions of each angles (0, 45, 90….360) are determined to find the actual radius and center (*xc*, *yc*) of each iris, as each angle consists of several points each point is considered the center and depending on the radius specified previously we draw a circle for each point to calculate the proportion of black and white after you take the largest percentage of blackness of the pixel density achieved by that angle and then calculate the radius and center for that angle to segmented the iris. This is the principle of Circular Distribution of Angles. Figure (9),(10) illustrated the circular distribution of the angles and table (1) , (2) describe the information of each angle and choice the angle that find the best center and radius of iris.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

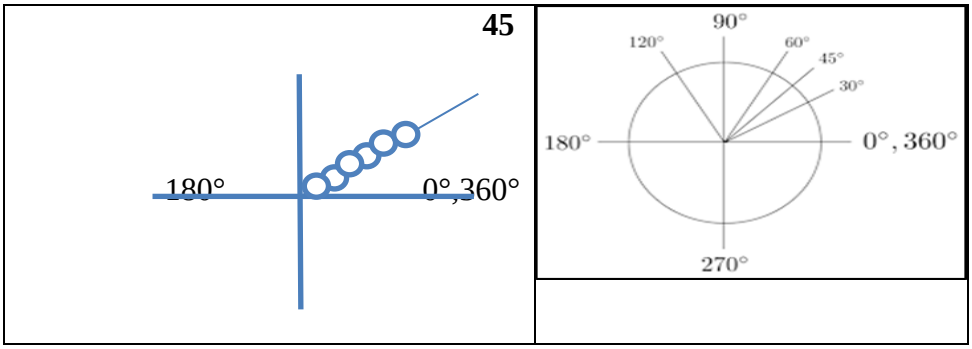


Figure (9): Circular distribution of angles ,(a) calculate circles in each angles , (b) directions of the angles

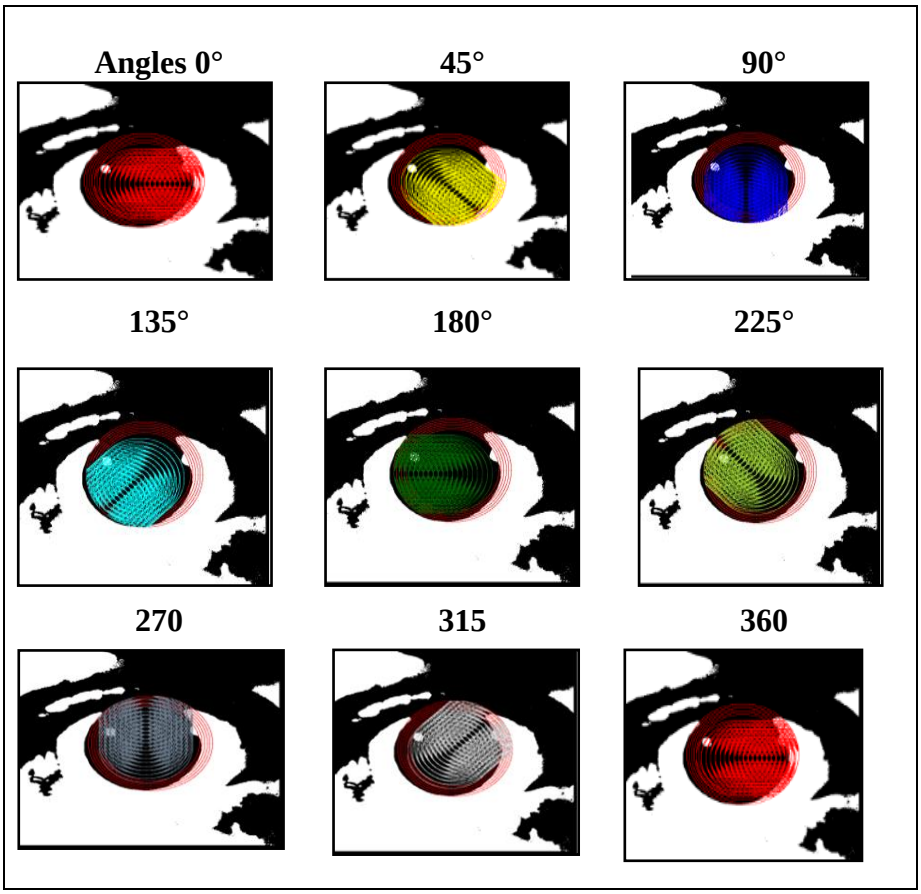


Figure 10: Circular distribution of angles

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

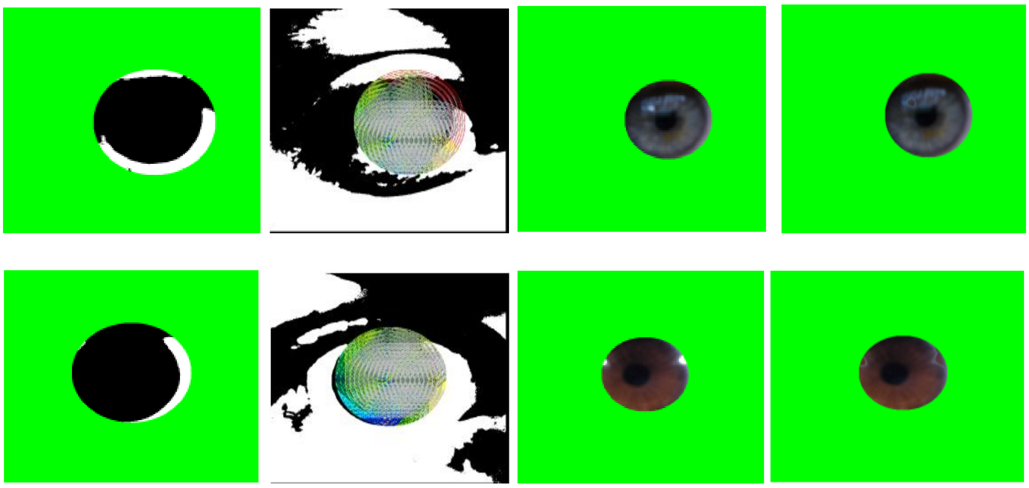
Table (1) Describe the information of each angles

	Center	0	45	90	135	180	225	270	315	360	405
▶	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	100	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0

Table (2) Example of the best angle that determine the fit of boundary of iris

	XC	YC	Angle Value	Radius	Angle
▶	309	258	100	90	0
*					

Hence, localizing the iris by drawing a perfect geometry that fits the boundaries as shown in figure (11).



(a) (b) (c) (d)

Figure (11): Iris Segmentation Output,(a) Remove interpolation, (b) Circular distribution of angles ,(c) Segmented iris (d) Remove interpolation after segmented iris.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

MICHE dataset

The MICHE dataset- created by BIP Lab of the University of Salerno- contains iris acquired using three different mobile devices: iPhone 5S, Samsung Galaxy S4, and a Samsung Galaxy Tab II. The images in the database are captured by the subjects holding the capturing device by herself/ himself [10]. The subjects were asked to capture their own iris images using both front and rear camera of these devices in two different scenarios −indoor and outdoor. There is a minimum of 40 images per subject. The average capture distances for smartphones were 10 cm for the front camera and about 13 cm for the rear camera. For the Tab, this stand-off distance is about 5cm. Examples of images in MICHE database are given in Figure 12.

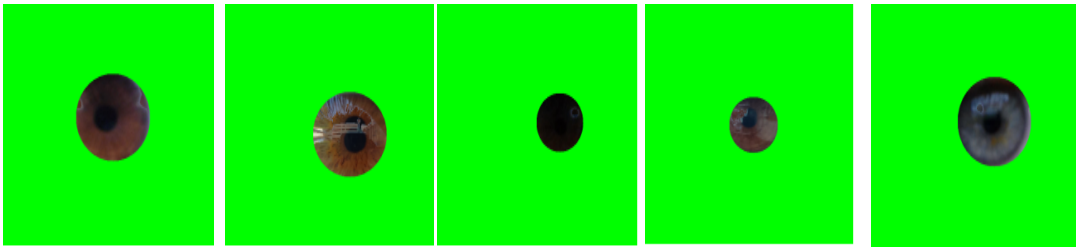
Figure 12: Examples of Images from the MICHE database: First row contains images captured using rear-facing camera and second row contains images captured using front-face camera**.**

Experimental results

The experiments have been conducted in unconstrained database (MICHE-I dataset (SG4)) , which includes 1297 images of irises captured from 75 person images containing either the left or the right iris, each person has been taken (8 images for training and 8 images for testing) in unconstrained outdoor and indoor environments with image size 520*520 pixel. In figure 13, we have chosen randomly a five images from the SG4 datasets in different environments. The results of our proposed method could localize the iris perfectly in unconstrained environment with high average accuracy rate 85% compared with the author Raja [5].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(a)



(b)

Figure 13: Results of iris segmentation on Galaxy Samsung S4 dataset, (a) Original images ,(b) iris segmented .

Conclusions

In this paper, a new method of image pre-processing in the smartphones environment is presented based on the circular histogram has helped us find initial centre in a database with difficult conditions and anew method of iris segmentation depended on circular distribution of the angles by finding the iris boundaries. The segmentation of the iris is very critical and affects the results of the iris recognition system. We investigated new method in order to improve the iris segmentation in unconstrained conditions through two stages , the first stage predicate the iris boundary. The second stage the distribution of angles (0, 45, 90 ... .360) are determined to find the actual radius and centre (xc,yc) for each iris. Our experiments have shown that circular distribution of angles was very robust in finding boundary of iris even in noisy and difficult conditions in smartphones environments .

The proposed approach system test and process have been implemented using a lap-top computer (processer: an Intel (R) Core (TM) i3 CPU M 380 @ 2.53 GHz with 32 bit Operating System and 4GB RAM) , the operating system is Windows 7. The programming language Visual C# is used to build and developed the required software.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1 shows a comparison of performance the proposed iris segmentation method with other existing methods using all iris images of MICHE-1 (Galaxy S4) dataset , the proposed method yields (85%) average accuracy rate. Even though all mobile devices are equipped with various biometric sensors, such as cameras, microphones, and fingerprint sensors, more research is needed in order to robustly consolidate the different biometric modalities on a single mobile device. Multimodal biometrics promises to be the future of mobile authentication more secure

Table 1: Calculation of Average Accuracy rate

Reference of method	**Av. Accuracy rate**
Kiran B.Raja[5] in SG4	Samsung Outdoor Rear 74.5 Samsung Outdoor Frontal 62 Samsung Indoor Rear 65 Samsung Indoor Frontal 77 69.62 %
Proposed method in SG4	**80 %**

References: -

[1] Thavalengal Shejin ,*" Contributions To Practical Iris Biometrics on Smartphones"*, Ph.D Thesis , National University of Ireland, Galway , College of Engineering and Informatics, 2016.

[2] Sinjini Mitra et al. "Biometrics in a Data Driven World: Trends, Technologies, and Challenges", chapter 2, 2017.

[3] S. Prabhakar, S. Pankanti, and A. Jain, "Biometric recognition: security and privacy concerns," IEEE Security Privacy, vol. 1, pp. 33–42, Mar 2003.

[4] Raghavender Reddy Jillela, Arun Ross," Segmenting Iris Images in the Visible Spectrum with Applications in Mobile Biometrics", Pattern Recognition Letters (2014).

[5] Kiran B. Raja , R. Raghavendra, Vinay Krishna Vemuri, Christoph Busch," Smartphone based visible iris recognition using deep sparse filtering", Pattern Recognition Letters (2014).

[6] A. F. Abate et al. , " BIRD : Watershed Based Iris Detection for mobile devices " , Pattern Recognition Letters (2014).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] S. Memar Zadeh and A. Harimi," Iris localization by means of adaptive thresholding and Circular Hough Transform", Journal of AI and Data Mining ,Vol 5, No 1, 21-28,2017.

[8] Narsi Reddy et al. " A Robust Scheme for Iris Segmentation in mobile environment" , IEEE 2016.

[9] A. Radman et al., "Automated segmentation of iris images acquired in an unconstrained environment using HOG-SVM and Grow Cut " , Digit. Signal Process. (2017).

[10] M. D. Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols," Pattern Recognition Letters, vol. 57, pp. 17 – 23, 2015.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Feature-Relationship Models: A Paradigm for Cross-hierarchy Business Constraints in SPL

Amougou Ngoumou, *Department of Computer Science, College of Technology,*
University of Douala,
PO 8698 Douala, Cameroon
ngoumoua@yahoo.fr
Marcel Fouda Ndjodo, *Department of Computer Science, Higher Teacher Training College,*
University of Yaounde I,
PO 47 Yaounde, Cameroon
marcel.fouda@dite-ens.cm

***Abstract*-In addition to the parental relationship between features, it appears in the scientific community that important cross-tree relationships exist between features in Software Product Line (SPL). Frequently, these relations are represented within SPL diagram such as feature diagram and this kind of representation can made diagrams become dense. Another representation using a separated graph near the feature diagram for each feature has been proposed. This representation gives a partial view of feature relations in a feature model. In this paper, a feature-association model for feature relations is proposed to express cross-hierarchy business constraints in software product lines. Since parental relationship between features in traditional feature model, don't capture the totality of constraints, the purpose of this work is to guarantee the validity of a selected product in a product line and to help stakeholders in the product choice process.**

I. INTRODUCTION

In the software engineering state of art, the mass customization of software products is known as software product lines [1] or software product families [2]. In order to achieve customer's personalization, software product line engineering promotes the production of a family of software products from common features instead of producing them one by one from scratch. This is the key change: software product line engineering is about producing families of similar systems rather than the production of individual systems.

Feature model languages are a common family of visual languages to represent software product lines [3]. The first formulation of a feature model language is due by Kang et al. in 1990 [4]. A feature model captures software product line information about common and variant features of the software product line at different levels of abstraction. A feature model is represented as a hierarchically arranged set of features with different relationships among those features. It models all possible products of a software product line in a given context. Unlike traditional information models, feature models not only represent a single product but a family of them in the same model.

To systematize the production of business components (i.e feature business components, subsystem business components, process business components and module business components), feature models have been extended to feature business components (FBC) [5, 6, 7, 8, 9, 10]. For this systematic production, the analysis of feature business components is essential since feature business components are the first assets produced in the chain. It is also usefull for a concrete software adaptation to requirements evolution and/or change.

Despite general properties, a feature business component can also contains cross-tree constraints between features. Those constraints are to be express when modeling a product line and have to be respected when one commands a specific product in the product line. To doing this by preserving coherence, it is essential to have a rigorous representation of cross-tree constraints. In this article, our aim is to achieve that goal.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The rest of the paper is organized as follow. In section 2 we present Feature business components, in section 3 we give general properties of feature business components, the cross-hierarchy business constraints are described in section 4, section 5 is dedicated to related works while section 6 draws conclusions and future research issues.

II. FEATURE BUSINESS COMPONENTS

In FORM, a feature model of a domain gives the "intention" of that domain in terms of generic features which literally mark a distinct service, operation or function visible by users and application developers of the domain. FORM/BCS specifies a feature model of a domain as a business reusable component of that domain which captures the commonalities and differences of applications in that domain in terms of features. Feature business components are used to support both the engineering of reusable domain artifacts and the development of applications using domain artifacts.

TABLE I
SPECIFICATION OF FEATURE BUSINESS COMPONENTS

FeatureBusinessComponent = = [name :**Name** ; descriptor:**Descriptor**; realization: **Realization** |

∀ fbc:**FeatureBusinessComponent,**
(solution(realization(fbc)) ∈ **Feature** ∧
adaptationpoints(realization(fbc)) ∈ 𝔽 (**Feature** × 𝔽 **Feature)**]
Feature = =[activity: **BusinessActivity** ; objects: **BusinessObjects** ;
relationships:[mandatory: 𝔽 **Feature**; optional: 𝔽 **Feature**; alternative: 𝔽 𝔽 **Feature**; or: 𝔽 𝔽 **Feature**] |]

In the above schemas, the type Feature specifies business activities. A business activity is caused by an event which is applied to a target set of objects. Features have relationships. Feature's relationships determines the set of (sub) features which have a mandatory-relationship with it and which indicate reuse opportunity, the set of (sub) features which have an optional-relationship with it, the set of groups of (sub) features which have an alternative-relationship with it, the set of groups of (sub) features which have an or-relationship with it.

A reusable feature business component fbc is well formed if it satisfies the following four characteristic properties which require that the realization section of a feature business component corresponds to the intention of that business component:

(fbc1) The solution given in the realization section of fbc is a solution of the intended contextual business activity of fbc:

action(domain(context(descriptor(fbc)))) = activity(solution(realization(fbc)))

(fbc2) The target of the intended contextual business activity of fbc is exactly the set of objects collaborating in the business activity of the solution given in the realization section of fbc:

target(domain(context(descriptor(fbc)))) = objects(solution(realization(fbc)))

(fbc3) Any requirement expressed in the form of a business process in the intended contextual business activity of fbc has a unique solution in the realization section of fbc:

∀ p∈ process(context(descriptor(fbc))),
∃ ! g ∈ relationships(solution(realization(fbc))) •
activity(g) = action(domain(p)) ∧ objects(g) = target(domain(p))

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(fbc4) Any solution in the relationships of the solution of the realization of fbc expressed in the form of a feature resolves a unique requirement expressed in the form of a business process in the intended contextual business activity of fbc:

$$\forall\, g \in \text{relationships}(\text{solution}(\text{realization}(\text{fbc}))),$$
$$\exists\,!\, p \in \text{process}(\text{context}(\text{descriptor}(\text{fbc}))) \bullet$$
$$\text{activity}(g) = \text{action}(\text{domain}(p)) \wedge \text{objects}(g) = \text{target}(\text{domain}(p))$$

A. Descriptors

The descriptor of a reusable business component gives an answer to the following question: "when and why use this component ?". A descriptor has an intention and a context. The intention is the expression of the generic modeling problem; the term "generic" here means that this problem does not refer to the context in which it is supposed to be solved. The context of a reusable business component is the knowledge which explains the choice of one alternative and not the other. Formally, descriptors are defined by the following schemas:

TABLE II
SPECIFICATION OF BUSINESS COMPONENT DESCRIPTORS

Descriptor = = [intention : **Intention** ; context : **Context** |]

Intention = = [action: **EnginneeringActivity**; target: **Interest** |]
Context = = [domain :**Domain** ; process : 𝔽 **Context** |]
EngineeringActivity = = **AnalysisActivity** | **DesignActivity**
AnalysisActivity = {analyze, ...}
DesignActivity = {design, decompose, describe, specify ,...}

The detailed specification is given in [5]. For the intelligibility of this paper, we give below an important type used in the above specification: Interest.

The engineering activity defined in the intention (hereafter referred to as the action of the intention) of a reusable business component acts on a "target" which can be a business domain or a set of business objects. Here are two examples of intentions formalized in FORM/BCS:

- $(\text{analyze})_{\text{ACTION}}(\text{civil servant management system})_{\text{TARGET}}$
- $(\text{describe})_{\text{ACTION}}(\text{civil servant recruitment application})_{\text{TARGET}}$

Interests of engineering activities are specified by the following schemas in which $\mathbb{F}\, A$ denotes the set of finite subsets of *A*.

TABLE III
SPECIFICATION OF INTENTION'S INTERESTS

Interest = **Domain** | **BusinessObjects**
Domain = = [action: **BusinessActivity**; target : **BusinessObjects** ;
precision : **Precision** |]
BusinessObjects = = 𝔽 **Class**
Class = = [name: **Name**; attributes : 𝔽 **Attribut**; operations : 𝔽 **BusinessActivity**|]
Precision
Name
Attribute

A business activity maintains relationships with a set of (sub) business activities divided into four disjoint categories: the set of mandatory (sub) business activities of the activity which indicate reuse opportunity (the commonality of the business activity), the set of optional (sub) business activities of the activity (the options of the business activity), the set of groups of alternative (sub) business activities of the activity (the switch ability of the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

business activity) and, the set of groups of inclusive (sub) business activities of the activity (the inclusive ability of the business activity). The capacity to have options, the switch ability and the inclusive ability for a business activity constitute its variability. A business activity is primitive (i.e. it has no relationship) or not.

TABLE IV
SPECIFICATION OF BUSINESS ACTIVITIES

BusinessActivity == [name: **Name**; mandatory: $\mathbb{F}$ **BusinessActivity** ; optional: $\mathbb{F}$ **BusinessActivity**; alternative: $\mathbb{F}$ $\mathbb{F}$ **BusinessActivity**; or: $\mathbb{F}$ $\mathbb{F}$ **BusinessActivity**; primitive: **Logic**|]

When the context is clear, given a business activity a, we write:

relationships(a) for mandatory(a) $\cup$ optional(a) $\cup$ ($\cup$(S $\in$ alternative(a))) $\cup$ ($\cup$(S $\in$ or(a))).

B. Realizations

The realization section of a reusable component provides a *solution* to the modeling problem expressed in the descriptor section of the component. It is a conceptual diagram or a fragment of an engineering method expressed in the form of a system decomposition, an activity organization or an object description. The goals, the activities and the objects figuring in the realization section concern the application field (product fragment) or the engineering process (process fragment).

The solution, which is the reusable part of the component, provides a product or a process fragment. The types of solutions depend on the type of reusable business component i.e a solution of a feature business component (respectively a reference business component) is a feature (respectively a reference business architecture). This solution may have *adaptation points* with values fixed at the reuse moment. Adaptation points enable the introduction of parameters in the solutions provided by reusable components. Those parameters are values or domains of values of elements of the solution.

TABLE V
SPECIFICATION OF BUSINESS COMPONENT REALIZATIONS

Realization = = [solution: **Solution**; adaptationpoints : **AdaptationPoints** |]
Solution = = **Feature**
AdaptationPoints = = $\mathbb{F}$ (**Feature** × $\mathbb{F}$ **Feature**)

III. CROSS-HIERARCHY BUSINESS CONSTRAINTS

Solutions of feature business components are represented using feature models. In basic feature models, there are two kinds of relationships between features [11]: parental relationships and cross-tree constraints between features. Concerning parental relationships, the following relations among features can exist:

- **Mandatory**: A child feature has a mandatory relationship with its parent when the child is included in all products in which its parent feature appears. For instance, every mobile phone system in the example (Figure 1) must provide support for *calls*.
- **Optional**: A child feature has an optional relationship with its parent when the child can be optionally included in all products in which its parent feature appears. In the example, software for mobile phones may optionally include support for *GPS*.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- **Alternative**: A set of child features have an alternative relationship with their parent when only one feature of the children can be selected when its parent feature is part of the product. In the example, mobile phones may include support for a *basic*, *colour* or *high resolution* screen but only one of them.
- **Or**: A set of child features have an or-relationship with their parent when one or more of them can be included in the products in which its parent feature appears. In Figure 1, whenever Media is selected, Camera, MP3 or both can be selected.

Notice that a child feature can only appear in a product if its parent feature does. The root feature is a part of all the products within the software product line.

Concerning cross-tree constraints between features which are our interest centre in this paper, two kinds of such constraints can be observed in feature models: hards constraints and soft constraints. Hard constraints express strong dependencies between features. Following the Mobile Phone example in Figure 1, we cannot have the Camera feature without the High resolution screen feature. Also, we cannot have a basic screen with the GPS feature. The explicit formalization of such constraints assures the assembling of valid feature combinations that will violate neither structural nor semantic dependencies of the product variant that has to be created. Therefore a given FM describes the set of valid combinations of features. Each combination, referred to as configuration corresponds to a specific product.

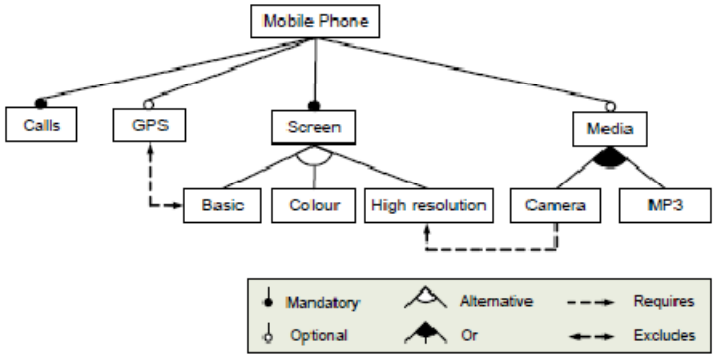


Figure. 1: Example of Mobile Phone Feature Model. Credits to Benavides et al. [citer Benavides]

Propositional formulas could be use to express feature models as presented in [12], [13]. The “inclusion” implication constraints are expressed with the requires statement and “exclusion” constructions are expressed with the excludes statement:

Examples:

- requires: $A \rightarrow B$ that is to say Feature A requires the presence of feature B.
- excludes: $A \rightarrow \neg B$ that is to say Feature A excludes the presence of feature B.

It is important to observe that, the expression of feature models as propositional formulas allows expressing more complex hards constraints relationship in the form of generic propositional formulas, for example, "A and B implies not C".

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Notice also that, because of logic inference rules, we can have hard constraints that are not explicitly formalized but that exist because of the other formalized constraints. This relation type between features is called inferred constraint:

Example: Camera $\rightarrow$ ⅂ Basic is an inferred constraint.

Soft constraints are motivated by the fact that the presence of two features in a given product configuration can also be qualified in terms of suitability. Such relations between features have also been explored in the literature. Bühne et al. [14] refer to them as hints and hinders relations. In the Pure::Variants commercial tool [15], the terms encourage and discourage are used in reference to the constraints in such relations. Czarnecki et al. [16] formally refer to them as two particular types of soft constraints, in the sense that their violation does not give raise to incorrect configurations. They exist with the objective of alerting the user, providing suggestions during the configuration process [16] or as a way to capture domain knowledge related to these trends. The notation below could be use for soft constraints:

Examples:
encourages: *soft*(A $\rightarrow$ B) that is to say Feature A encourages the presence of feature B.
discourages: *soft*(A $\rightarrow$ ⅂B) that is to say Feature A discourages the presence of feature B.

This way, a domain expert of the Mobile Phone example could explicitly formalize the soft constraint *soft (GPS $\rightarrow$ High resolution)* mentioned before to capture this SPL domain knowledge.

In an analogous manner, it is also important to observe that, the expression of feature models as propositional formulas allows expressing more complex soft constraints relationships in the form of generic propositional formulas, for example, "A and B encourage C" or "A and B discourage C".

IV. Cross-hierarchy business constraints capture with feature-relationships models

In this section, we give first the theoretical foundation of feature-relationship models, then we present feature-relationship diagrams and thirdly, a case study is done to show how feature-relationships models align with cross-hierarchy business constraints.

A. Feature-relationships models

A feature-relationship model represents the information of all cross-hierarchy business constraints of a software product line in terms of features and non parental relationships among them. Feature-relationship models are inspired from entity-relationship models which are conceptual kind of datas models. Like thus, they are actually used by many methods and datas base dessign aid tools (MERISE, IDA, Yourdon, …). A feature-relationship model is a desciption based on three basic concepts which are features, cross-hierarchy business relationships between features and properties of those features.

Features: A feature is a prominent or distinctive and apparent aspect, quality, or characteristic of a system [4]. In feature-relationship models, we capture a feature by given its name, business activity and the list of its business

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

objects. Features are represented by rectangular, business activities and business objects are listed inside the rectangular representing the feature. The graphical notation below is used to represent features.

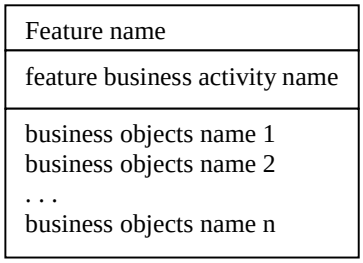


Figure 2: Feature notation

Cross-hierarchy business relationships: A cross-hierarchy business relationships is a non parental relationship between features. We distinguish four kinds of cross-hierarchy business relationships :

- **Simple hard constraints relationships** between features in a software product line that are typically inclusion and exclusion statements in the form : if feature F is included, then feature A must also be included (or excluded).

Graphically, a simple hard constraint relationship is modeled folowing the notation below :

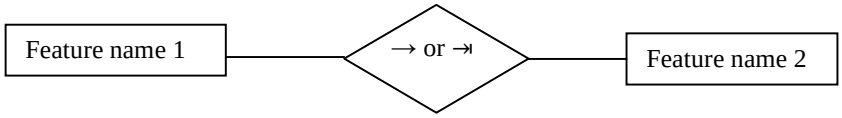


Figure 3: Simple hard constraint relationship between features

In figure 3 feature's notation is abridged using only the name part. the arrow in the lozenge is written to mean requires or excludes. The Arrow indicates the sense of the requirement. Thus, it means here that Feature1 requires Feature 2 (or Feature1 excludes Feature 2).

- **Complex hard constraints relationships** between features in a software product line are capture in the form of generic propositional formulas, e.g. "A and B implies not C". Someone could observe that, these kinds of constraints between features may involve many features more than only three features.

Graphically, a complex hard constraint relationship is of course modeled using complex notations like those in figure 4.

Figure 4: Feature relationship notations

Let us consider the following hard complex constraint:

"(A and B and C and (D xor E) and (F or G)) implies H"

This constraint is graphically represented as follows:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

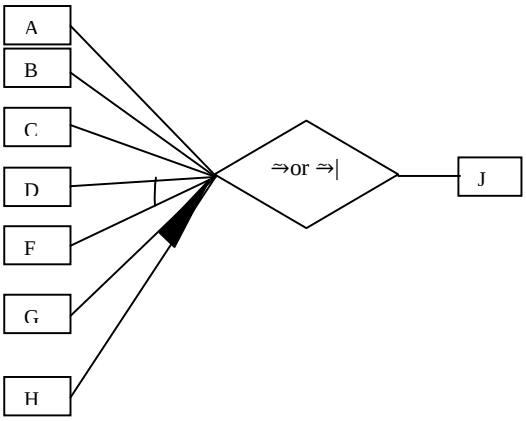


Figure 9: Left side complex soft constraint relationship between features

We could also imagine the following hard complex constraint in which the complexity is situated in the right side: *H implies (A and B and C and (D xor E) and (F or G)).* This constraint could be represented as follows:

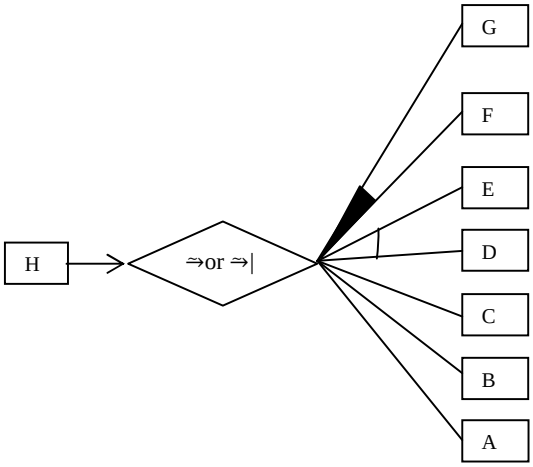


Figure 10: Right side complex soft constraint relationship between features

Finally, we could encounter the following hard complex constraint in which the complexity is situated in the both sides:

A and B and C and (D xor E) and (F or G)) implies (A and B and C and (D xor E) and (F or G)).

This constraint could be represented as follows:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

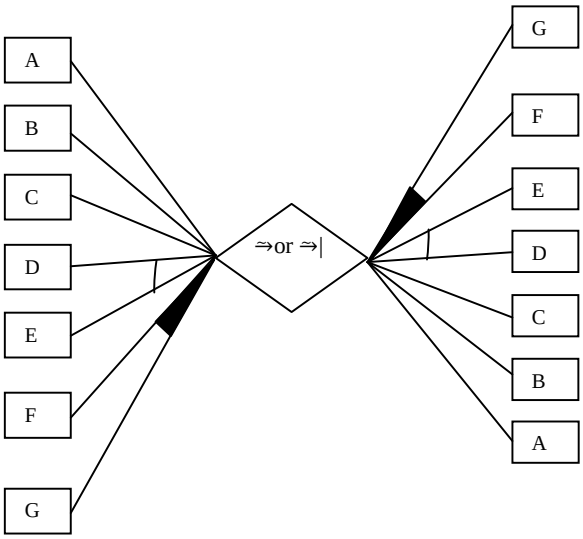


Figure 11: Both sides complex soft constraint relationship between features

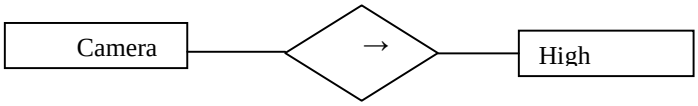
B. Feature-relationship diagrams

The feature-relationship model allows a graphical representation of all cross-hierarchy business constraints of a software product line called feature relationship diagram.

When the analysis of a product line is performed, this diagram is separated from the feature business component and not in the same component to avoid it to be crowded and complex. It completes the understanding of the solution part (which is a feature model) of this service perspective with cross-hierarchy business constraints.

To illustrate what we are saying in a simple manner, let us take the example of Mobile Phone Feature Model given in figure 1. In this model, there are two cross-hierarchy business constraints: (1) we cannot have the Camera feature without the High resolution screen feature; (2) we cannot have a basic screen with the GPS feature. To express theses constraints in the new approach, instead of putting cross-hierarchy business constraints in the model, they are represented outside the model. Below, Figure 12a represents the cross-hierarchy business constraint between the camera feature and the high resolution screen feature while Figure 12b represents the cross-hierarchy business constraint between the basic screen feature and the GPS feature.

(a)



(b)

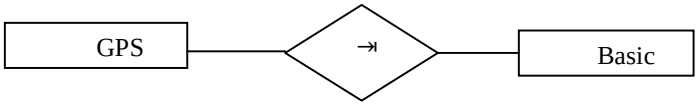


Figure 12: Cross-hierarchy business constraints

By representing cross hierarchy business constraints outside the feature business component, there are four benefits:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- it is possible to represent graphically complex cross-hierarchy business constraints;
- the separation of preoccupations in the two representations allows to concentrate effort each time in only one aspect;
- the possibility to handle separately the feature business component without taking cross-hierarchy business constraint into consideration is offered and vice versa;
- we have the possibility to represent all cross-hierarchy business constraints in complex systems without making the model become unreadable.

Feature-relationship diagrams enable to define the following functions:

- *requiringfeatures* which, given a feature, provides the set of features requiring it;
- *excludingfeatures* which, given a feature, provides the set of features excluding it;
- *featureweight* which, given a feature, provides the number of features which require it;
- *featurelightness* which, given a feature, provides the number of features which exclude it;

featureweight: **Feature ↔ Integer**
$f \rightarrow \#(requiringfeatures(f))$

featurelightness: **Feature ↔ Integer**
$f \rightarrow \#(excludingfeatures(f))$

C. The case study

To illustrate cross hierarchy business constraints in SPL, we take the sample of biometric lock. This device enables the control of fingerprints for each person wanting to enter an entry. After the analysis of this device, we have found the following features:

- *administration*: this feature enables the enrolment of fingerprints and the change of battery;
- *authorization*: this feature enables the reading of fingerprints, the warning of the results of fingerprints control and the unlock of the biometric lock;
- *signalization*: this feature enables the indication of the battery charge level and the resistance again intrusions.

If we use a feature diagram to model the biometric lock, it will look like the following figure:

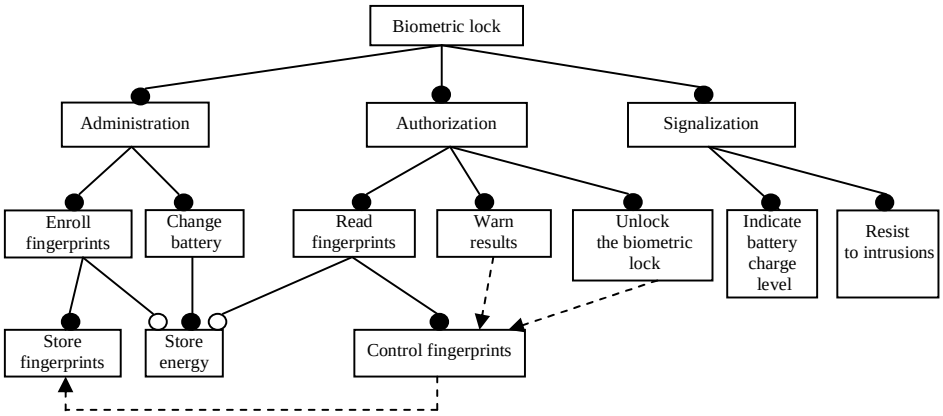


Figure 13: Feature diagram for biometric lock

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The require relations between feature makes the diagram become dense. For example the feature Control fingerprints has three require relations with the feature Warn results, Unlock the biometric lock and Store fingerprints. These relations can be clearly represent in a feature-association separated model as follows

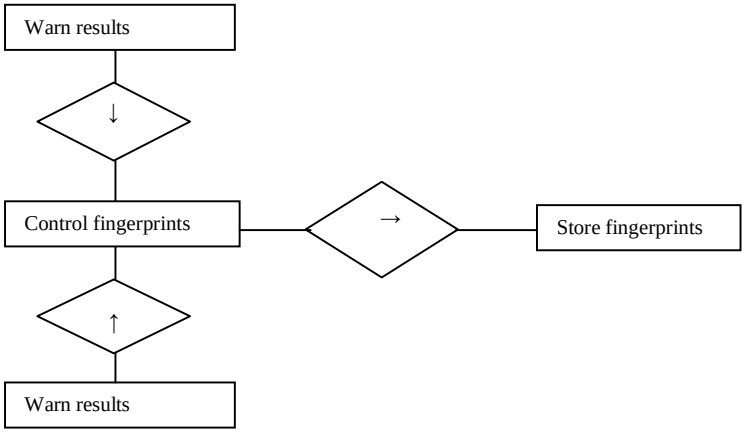


Figure 14: The Control fingerprints with three *require* relations

This representation shows clearly the features for which the weight or the lightness is not equal to zero. Thus:

weight (Control fingerprints) = 2
weight (Store fingerprints) = 1

By removing cross hierarchy relation in the feature diagram, it becomes easily understandable. The biometric lock feature diagram in which the cross hierarchy relations has been removed is the following.

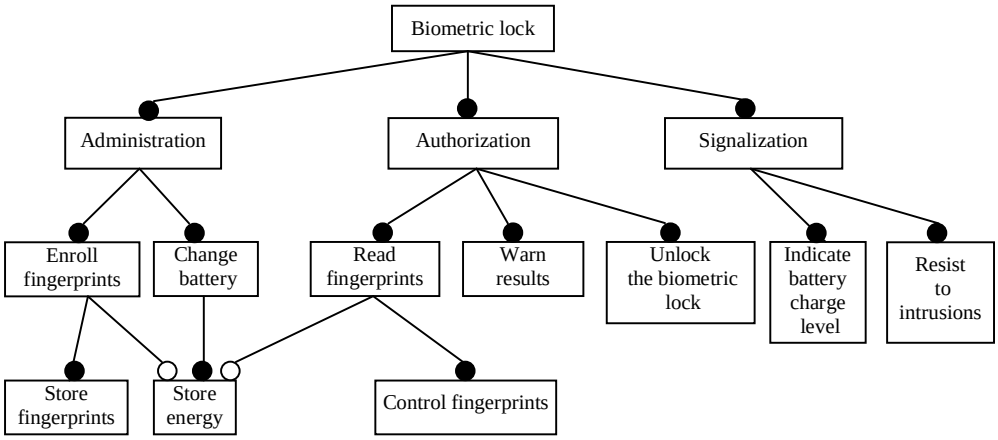


Figure 15: The biometric lock feature diagram without cross hierarchy relations

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. Related works

A feature model represents the information of all possible products of a software product line in terms of features and relationships among them [17]. Feature models are a special type of information model widely used in software product line engineering. A feature model is represented as a hierarchically arranged set of features composed by:

1. relationships between a parent (or compound) feature and its child features (or subfeatures).
2. cross–tree (or cross–hierarchy) constraints that are typically inclusion or exclusion statements in the form: if feature F is included, then features A and B must also be included (or excluded).

Feature Relations Graphs (FRoGs) have been proposed in [11]. This visualisation paradigm, called FRoGs, has as objective to visualise the constraints (relations) between features based on the analysis of current configurations. Each FRoG is associated with a specific feature to show how this feature is related, in term of hard and soft constraints, to the other features. A FRoG is presented as a circle where the considered feature is displayed in the center. This feature in the center will be called f_c hereafter. All the rest of features that represent variation points of the FM are displayed around f_c with a constant separation of 2π / (|features|−1). This separation allows to distribute uniformly all the features around the circle. These features are ordered by stakeholder perspectives so we will have circular sectors for each of the stakeholder perspectives.

VI. Conclusions and Future research

In this paper, a feature-association model for feature relations has been proposed to express cross-hierarchy business constraints in software product lines. Since parental relationship between features in traditional feature model, don't capture the totality of constraints, the purpose of this work is to guarantee the validity of a selected product in a product line and to help stakeholders in the product choice process.

After the enrichment of feature models with cross hierarchy business constraints, we plan to elaborate a suitable architecture for a framework supporting the feature oriented reuse method with business component semantics.

References

[1] P. Clements and L. Northrop. Software Product Lines: Practices and Patterns. SEI Series in Software Engineering. Addison–Wesley, August 2001.
[2] K. Pohl, G. Bockle and F. van der Linden. Software Product Line Engineering: Foundations, Principles, and Techniques. Springer–Verlag, 2005.
[3] P. Schobbens, J.C. Trigaux P. Heymans, and Y. Bontemps. Generic semantics of feature diagrams. Computer Networks, 51(2):456–479, Feb 2007.
[4] K. Kang, S. Cohen, J. Hess, W. Novak, and S. Peterson. Feature–Oriented Domain Analysis (FODA) Feasibility Study. Technical Report CMU/SEI-90-TR-21, Software Engineering Institute, Carnegie Mellon University, November 1990.
[5] M. Fouda, N. Amougou, "The Feature Oriented Reuse Method with Business Component Semantics", International Journal of Computer Science and Applications, Vol. 6, No. 4, pp 63-83, 2009.
[6] M. Fouda, N. Amougou, "Product Lines' Feature-Oriented Engineering for Reuse: A Formal Approach", International Journal of Computer Science Issues, Vol. 7, Issue 5, pp 382-393, 2010.
[7] N. Amougou, "Extension de la méthode FORM pour la Production des Architectures Adaptables de Domaine", PhD thesis, University of Yaounde I, Yaounde, Cameroon, 2011.
[8] M. Fouda, N. Amougou, "Transformational Variability Modelling Approach To Configurable Business System Application", in Software Product Line – Advanced Topic, Edited A. O. Elfaki, Intech Publisher, pp. 43-68, 2012.
[9] N. Amougou, M. Fouda, , "A Rewriting System Based Operational Semantics for the Feature Oriented Reuse Method", International Journal of Software Engineering and Its Applications, Vol.7, No.6, pp.41-60, 2013.
[10] N. Amougou, M. Fouda, , "A Command Oriented Derivation Approach with Product-Specific Architecture Optimization", International Journal of Software Engineering and Its Applications, Vol. 9, No. 2, pp. 23-40, 2015.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[11] J. Martinez, T. Ziadi, R. Mazo, T. F. Bissyandé, J. Klein, et al.. Feature Relations Graphs: A Visualisation Paradigm for Feature Constraints in Software Product Lines. IEEE Working Conference on Software Visualization (VISSOFT 2014), Sep 2014, Victoria, Canada. pp.50 - 59, 2014.
[12] D. S. Batory. Feature models, grammars, and propositional formulas. In J. H. Obbink and K. Pohl, editors, SPLC, volume 3714 of Lecture Notes in Computer Science, pages 7–20. Springer, 2005.
[13] K. Czarnecki and A. Wasowski. Feature diagrams and logics: There and back again. In SPLC, pages 23–34. IEEE Computer Society, 2007.
[14] S. Bühne, G. Halmans, and K. Pohl. Modeling dependencies between variation points in use case diagrams. In Proceedings of 9th intl. Workshop on Requirements Engineering - Foundations for Software Quality, pages 59–69, 2003.
[15] D. Beuche. Modeling and building software product lines with pure::variants. In SPLC Workshops, page 296, 2010.
[16] K. Czarnecki, S. She, and A. Wasowski. Sample spaces and feature models: There and back again. In SPLC, pages 22–31. IEEE Computer Society, 2008.
[17] D. Benavides. On the Automated Analyisis of Software Product Lines using Feature Models. A Framework for Developing Automated Tool Support. PhD thesis, University of Seville, 2007.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TITLE: A Clustering Based Approach To End Price Prediction In Online Auctions

Author 1:

Dr. Preetinder Kaur,
Course Convenor – ICT,
Western Sydney University International College, Sydney
100 George Street, Parramatta NSW 2150

Email: preetinder77@yahoo.com

Author 2:

Dr. Madhu Goyal
Feculty of Engineering and Information Technology,
University of Technology, Sydney
15 Broadway, Ultimo NSW 2007

Email: madhu.goyal-2@uts.edu.au

Corresponding author:

Dr. Preetinder Kaur,
Course Convenor – ICT,
Western Sydney University International College, Sydney
100 George Street, Parramatta NSW 2150

Email: preetinder77@yahoo.com
Phone: +61 0401210451

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Clustering Based Approach To End Price Prediction In Online Auctions

Abstract- Online auctions have emerged as a well-recognised paradigm of item exchange over the past few years. The availability of numerous auctions of similar items complicates the bidders' situation for choosing the auction where their participation will give maximum surplus. Further, the diverse price dynamics of auctions for same or similar items affect both the choice of auction and the valuation of the auctioned items. This paper presents a design of end price prediction methodology which finds an auction to compete in and assesses the value of the auctioned item using data mining techniques. It handles the problem of diverse price dynamics in auctions for same or similar items, using a clustering-based bid mapping and selection approach to find the auction with maximum surplus. The proposed approach is validated using real online auction datasets. These results demonstrate that this clustering-based price prediction approach outperforms existing methodologies in terms of prediction accuracy.

Keywords: Data Mining, Clustering, Online Auctions, Price Prediction, Multiple linear regression, K-nearest neighbour, eBay, Software agents, E-commerce, Bid mapping .

1. Introduction

The advent of electronic commerce has dramatically advanced traditional trading mechanisms, and online auction settings like eBay and Amazon have emerged a powerful tool for allocating goods and resources. Discovery of the new markets and the possibilities opened by online trading have heightened the sellers' and buyers' interest. In recent years, online auctions have become a widely recognised paradigm of item exchange, offering traders greater flexibility in terms of both time and geography. In online auction commerce, traders barter over products, applying specific trading rules over the Internet and supporting different auction formats. Common online auction formats are English, Dutch, First-price sealed-bid and Second-price sealed-bid auctions.(Haruvy & Popkowski Leszczyc, 2010; Ockenfels, Reiley, & Sadrieh, 2006)

Bidders in this marketplace often feel challenged when looking for the best bidding strategies to win the auction. Moreover, there are commonly many auctions selling the desired item at any one time. Deciding which auction to participate in, whether to bid early or late, and how much to bid are very complicated issues for bidders.(Wolfgang Jank & Zhang, 2011; Park & Bradlow, 2005) The difficult and time-consuming processes of analysing, selecting and making bids and monitoring developments need to be automated in order to assist buyers with their bidding.

The emergence of software agent technology has created an innovative framework for developing online auction mechanisms. Because of their extraordinary adaptive capabilities and trainability, software agents have become an integral component of online trading systems for buying and selling goods. These software agents represent expert bidders or sellers to fulfil their requirements and pursue their beliefs, and are

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

consequently trained to achieve these aims. Software agents can perform various tasks like analysing the current market to predict future trends, deciding bid amounts at a particular moment in time, evaluating different auction parameters and monitoring auction progress, as well as many more. These negotiating agents outperform their human counterparts because of the systematic approach they take to managing complex decision-making situations effectively.(Byde, Preist, & Jennings, 2002) This creates more opportunities for expert bidders and sellers to maximise satisfaction and profit. Software agents make decisions on behalf of the consumer and seek to guarantee that items are delivered to the buyer's preferences. To function well, these agents must have prior knowledge of the auction's features, whether these are certain or uncertain.

eBay is one of the major global online marketplaces and currently the biggest consumer-to-consumer online auction site. Founded in 1995, eBay Inc. has attracted over 112 million active users and touted net revenue of $14.1 billion for 2012[1]. eBay does not, however, actually sell any goods that it owns; it only makes the process of displaying and selling goods easier by facilitating the bidding and payment processes. In virtual terms, eBay provides a marketplace where buyers and sellers meet and transact. eBay is a great source of high quality data as it keeps detailed records of completed auctions, and this data has been used extensively by researchers to solve research issues involving online auctions.(Bajari & Hortacsu, 2003; Bapna, Goes, & Gupta, 2003; W Jank & Shmueli, 2005; Ilya Raykhel & Ventura, 2009; Roth & Ockenfels, 2002; Van Heijst, Potharst, & Van Wezel, 2008; Wilcox, 2000)

The eBay-style auctions adopt the English auction format, except with regard to the payment of the winning bid(Haruvy & Popkowski Leszczyc, 2010). In eBay auctions, the winner is the bidder with the highest value bid, but instead of paying his own bid, he pays the second-highest bid plus the amount of one bid increment. Bidders in these auctions do not, however, bid their maximum valuation of the item offered. This is either because they do not grasp that they should do so or they simply have trouble figuring out what their maximum valuation is. These bidders are typically afraid of winning the auction at a price above the true value of the item, a phenomenon sometimes known as 'the-winner's curse".(Ku, 2000) This problem occurs because the bidder lacks information about the true value of the item. In this respect, closing or end price predictions can assist bidders to assess the true value of the item on auction and thus finalise the maximum amount that they are willing to pay. This helps them to develop a bidding strategy to win the auction if the price is appropriate to an item's value, and it also allows experienced bidders to win auctions at a lower cost.(Sun, 2005) By presenting consistent information, end price prediction supports buyers to make more informed bidding decisions.(I Raykhel & Ventura, 2008) This also solves some of the information asymmetry problem for buyers, cutting down transaction times and costs. At the same time, sellers can use predictions to identify when the market favours selling their products and assess the value of their inventory better. They can also optimise auction attributes and the selling price for their wares.(Xuefeng, Lu, Lihua, & Zhao, 2006)

[1] eBay Annual Report 2012

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Furthermore, an item's value strongly depends on the path that its price takes during its auction, or what is known as the price dynamics of the auction. Therefore, a methodology is required that assesses the true value of items based on the price dynamics of their auctions. Currently the most widely used methods for auction selection and value assessment item are static; they use information available only at the beginning of an auction.(Ghani & Simmons, 2004; Van Heijst et al., 2008) These models cannot incorporate the price dynamics of auctions for similar items, a concern that only a few studies have pursued.(Dass, Jank, & Shmueli, 2011; Kehagias, Symeonidis, & Mitkas, 2005; Wang, Jank, & Shmueli, 2008) Further, the price dynamics of auctions are different, even when they deal with similar items, and this has a profound influence on the choice of an auction and valuation of the item being auctioned.(W Jank & Shmueli, 2005) We need, thus, to characterise these auctions based on their price dynamics before we can select one to participate in and assess the true value of the goods on offer.

Against this background, the research reported in this paper takes as its object a theory and methodology for predicting the end price of an auction in order to choose an auction to participate in from simultaneous auctions for the same or similar items and to assess the value of the auctioned item. To achieve these aims, a clustering approach is adopted along with bid mapping and selection technique. The prediction methodology predicts the amount of a bid in an ongoing auction at a particular moment during the auction and at the end of the auction. This contrasts with static prediction models that solely predict the closing price of the auction based on information which is available only at the start of the auction.(Ghani & Simmons, 2004; Van Heijst et al., 2008) The static models may use auction attributes such as the opening bid, the auction type, the length of the auction, the seller's rating and the product description. These attributes are decided at the beginning of an auction and remain fixed for its active period. At the same time, however, these models do not incorporate the dynamic attributes which are generated using information that arrives during the active period of the auction such as the bid rate of the auction, the number of auction participants and the amount of a bid at a particular time and moment of competition. The price prediction methodology presented in this paper therefore extracts these essential auction features from the auction history in order to assess the value of the item based on dynamic auction attributes.

The rest of the paper follows this schema: Section 2 designs the methodology for end price prediction in order to select a target auction to participate in and assess the value of its item. Section 3 describes the approaches adopted for validating the end price prediction methodology. Section 4 presents the experiments performed and the results achieved. Section 5 concludes the paper and presents the future work.

2. End price prediction

This paper proposes a design of a price prediction methodology which selects an auction to participate in from simultaneous auctions for the same or similar items and assesses the value of the auctioned item. The price is predicted using a clustering approach combined with a bid mapping and selection technique.(Kaur, Goyal, & Lu, 2012) The main purpose here is to use the price dynamics of different auctions both to choose a target auction to compete in and to assess the value of the auctioned item. The price

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

prediction approach consists formally of three steps: cluster analysis, bid mapping and selection, and value assessment. The first step aims to cluster similar auctions together in *k* groups based on their price dynamics. In the second step, the bid mapping and selection component nominates the cluster for each ongoing auction. A target auction is selected for participation based on the transformed data after clustering and the characteristics of current auctions. The last step applies machine learning algorithms to assess the value of the item in the selected auction.

2.1. *Cluster Analysis*

The price dynamics of online auctions are different, even when those auctions are for similar items.(W Jank & Shmueli, 2005) The value of an item assessed by the end price prediction methodology therefore depends on the path that its price follows during an auction. These varying price dynamics recommend partitioning of the input auction space into groups of similar auctions before we actually predict bid amounts—if we want to improve prediction outcomes. The price prediction methodology, thus, divides the auction space into a set of auctions with similar price dynamics before proceeding to forecast the closing price. In this paper, a clustering-based approach is responsible for partitioning auctions with similar price dynamics.

In the proposed methodology, historical auction data is extracted for the required dynamic attributes. These form the agent's knowledge base of auctions for cluster analysis. Let *ATT* be the set of attributes collected for each auction, so that $ATT= [a_1, a_2 \ldots\ldots a_j]$ where *j* is the total number of attributes. Different types of auctions are categorised based on the attributes of online auctions. These attributes may include the average bid amount, the average bid rate, the number of bids, the item type, the seller's reputation, the opening bid, the closing bid, the quantity of items available in the auction, the type of auction, the duration of the auction, the buyer's reputation and many more. To classify the different types of auctions, this paper concentrates on a single set of attributes; the opening bid, the closing price, the number of bids, the average bid amount and the average bid rate. Of these, the opening bid and closing price are static attributes, i.e. they do not shift over time, while the number of bids, the average bid amount and the average bid rate are dynamic attributes, which change as the auction progresses.

Now $ATT=[OpenB_i, CloseP_i, NUM_i, AvgB_i, AvgBR_i]$

where *ATT* is the set of attributes for an auction

$OpenB_i$ is the opening bid or the starting price of the i^{th} auction

$CloseP_i$ is the closing price of the i^{th} auction

NUM_i is the total number of bids placed in the i^{th} auction

$AvgB_i$ is the average bid amount of the i^{th} auction and can be calculated as $Avg(B_1, B_2, \ldots\ldots B_l)$ where B_1 is the 1st bid amount, B_2 is the second bid amount and B_l is the last bid amount for the i^{th} auction.

$AvgBR_i$ is the average bid rate of the i^{th} auction and is calculated as

$$AvgBR_i = \frac{1}{n}\sum_{i=1}^{n}\frac{B_{i+1}-B_i}{t_{i+1}-t_i} \quad (1)$$

where B_{i+1} is the amount of $(i+1)^{th}$ bid, B_i the amount of the i^{th} bid, t_{i+1} is the time at which $(i+1)^{th}$ bid is placed and t_i is the time at which the i^{th} bid is placed.

The input auctions are partitioned into groups of similar auctions depending on their different characteristics. This partitioning has been achieved by using the *k-means* clustering algorithm. The main objective of the *k-means* clustering is to define *k*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

centroids, one for each cluster. In this paper, the Elbow approach using one way analysis of variance (ANOVA) is explored to estimate the value of *k*.(Kaur et al., 2012)

Once we decide the value of *k*, the *k-means* clustering algorithm is used to partition the similar auctions based on their characteristics. Given a set *A* of *N* auctions $A=[a_1,a_2,...a_N\ ]$ where each auction is a *5-dimensional* real vector *ATT* and $ATT=[OpenB_i, CloseP_i, NUM_i, AvgB_i, AvgBR_i]$, *K-means* clustering aims to partition *N* auctions into *k* clusters *(k<N)* so that within-cluster dispersion is minimal. The within-cluster dispersion is the sum of squared Euclidean distances of auctions from their cluster centroid. The steps of the *k-means* clustering algorithm are as follows:

Initialisation step: Initialise *k*-centroid vectors for *k* initial clusters as $c_1, c_2, ...c_k$ where $c_n = \{OpenB_n, CloseP_n, NUM_n, AvgB_n, AvgBR_n\}$ and $OpenB_n$ is calculated as $(OpenB_{1n}+ OpenB_{2n} +.....+ OpenB_{Nn})/N$, where *N* is the number of auctions in the n^{th} cluster. The values of $CloseP_n$, NUM_n, $AvgB_n$, and $AvgBR_n$ are calculated similarly.

Assignment step: Assign each auction to a cluster with the closest mean by calculating the Euclidean distance of each auction from each of these *k* centroids as follows:

$$d_{in} = \sqrt{(OpenB_i - OpenB_n)^2 + (CloseP_i - CloseP_n)^2 + (NUM_i - NUM_n)^2 + (AvgB_i - AvgB_n)^2 + (AvgBR_i - AvgBR_n)^2} \quad (2)$$

where d_{in} is the Euclidean distance of the i^{th} auction from the n^{th} cluster.

Update step: Calculate the new means as *k*-centroid vectors for *k* auction clusters which are generated in the assignment step as $C_1, C_2, ...C_k$ where $C_n = \{OpenB_n, CloseP_n, NUM_n, AvgB_n, AvgBR_n\}$ and $OpenB_n$ is calculated as $(OpenB_{1n}+ OpenB_{2n} +.....+ OpenB_{Nn})/N$, where *N* is the number of auctions in the n^{th} cluster. The values of $CloseP_n$, NUM_n, $AvgB_n$, and $AvgBR_n$ are calculated similarly.

We repeat the assignment and update steps until moving an auction between clusters increases the within-cluster dispersion.

2.2. *Bid Mapping and Selection*

In order to decide the cluster which current ongoing auctions belong to, the bid mapping and selection component is activated. Based on the transformed post-clustering data and the current auctions' characteristics, the component nominates a cluster for each ongoing auction so that one can be chosen to participate in.

The use of a Bid Mapper and Selector (BMS) algorithm is proposed for the process of finding the target auction to compete in. The clustering technique described in the previous section characterises different auctions following a specific price path based on the distinct range of the average bid rate (*AvgBR*). This paper focuses on auctions with hard closing rules like eBay auctions. These hard closing rules persuade many bidders not to bid until the closing moments of the auction, so these bidders do not reveal the maximum they are willing to pay during the early moments of the auction. To neutralise this tendency, eBay uses a proxy bidding system which allows a bidder to submit the maximum amount that he is prepared to pay instead of his current bid. The proxy system keeps this amount private; it bids on the bidder's behalf at just one increment over the next highest bid until it reaches the maximum that the bidder is willing to hand over. This helps to reduce hikes in price paths near the end of the auction and to keep the average bid rate (*AvgBR*) at almost the same level in the closing moments as earlier on. So, the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

BMS algorithm uses the *AvgBR* value near the closing time when it maps the ongoing auctions to the clusters.

```
K ← total number of clusters
N ← total number of auctions
Cnt_k ← counts the number of auctions in the k^th cluster
AvgBROA_i ← average bid rate of the i^th ongoing auction
MinAvgBRC_k ← minimum average bid rate of the auctions in the k^th cluster
MaxAvgBRC_k ← maximum average bid rate of the auctions in the k^th cluster
OA_k←set of auctions in the k^th cluster
AvgBC_k ← average bid amount of the auctions in the k^th cluster
TA ← target auction for participation
AvgB_TA ← average bid amount of the target auction


AvgB_TA ← 0;
TA← 0;
for  i ← 1 to N {       // Check all the auctions.
        Cnt_k ←  0    //Init variable that counts the auctions in a cluster.
        for k ← 1 to K {       // Loop to check all the clusters.

                // Check if the value is in range of the cluster.
                if ( MinAvgBRC_k ≤ AvgBROA_i ≤ MaxAvgBRC_k ) {
                        OA_k ← OA_k U OA_i; // Add the auction to the current cluster.
                        Cnt_k = Cnt_k +1;       // Increment the cnt.

                        if (AvgBOA_i ≤ AvgBC_k){    //Compare with the cluster min.
                                AvgBC_k = AvgBOA_i; //New cluster min.
                        }//end if

                        if(AvgB_TA = 0 OR AvgBC_k < AvgB_TA ) {
                            AvgB_TA = AvgBC_k;
                            TA = OA_i;          // Target auction.
                        }//end if
                }// end if
        } // for 1 to K
} //for1 to N
```

Fig. 1. Bid Mapper and Selector algorithm (BMS)

Given a set of ongoing auctions $OA = OA_1 \cup OA_2 \cup \ldots\ldots \cup OA_k$. where $OA_i = [OA_{i1}, OA_{i2}, \ldots\ldots OA_{in}]$, OA_i is the set of ongoing auctions belonging to the i^{th} cluster , $i=1,2,\ldots.k$ where k is the total number of clusters and n is the total number of ongoing auctions belonging to the i^{th} cluster, the BMS algorithm selects a subset of OA if $AvgB_{kn} < AvgBC_k$, where $AvgB_{kn}$ is the average bid amount of the n^{th} auction in the k^{th} cluster and $AvgBC_k$ is the average bid amount of the k^{th} cluster.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The target auction *TA* is selected on the basis that it gives the maximum surplus to bidders. The surplus is the return that bidders enjoy by winning an auction at a price lower than its predicted closing price.(Dass et al., 2011) Auctions with lower average bid amounts are expected to give higher surplus to bidders. The target auction *TA* is, thus, selected for participation on the basis that its average bid amount is lowest in order to maximise bidders' surplus. Figure 1 presents the developed algorithm:

2.3. *Value Assessment*

Once an auction is selected for participation, the next task is to assess the value of the auctioned item. The value of the item is assessed by predicting the closing price of the auction. Knowing the closing price predicted for the auction helps bidders to establish the true value of the item and in turn finalise their own maximum valuation of the auctioned item. The closing price prediction task is handled using machine learning algorithms. Two approaches were considered in predicting the continuous price value of the item: parametric and non-parametric.

2.3.1. *Parametric Approach to Price Prediction*

Multiple linear regression is the most common parametric approach to making predictions. This model has been used to fit a linear relationship between the dependent variable (the closing price) and a set of predictor variables. The multiple linear regression model is employed based on the distinct clusters' characteristics. To predict the closing price of the selected auction, regression coefficients w_j for each attribute are opted such that the sum of squares between the predicted and the actual bid (3) over all the training auction data is minimal.

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{m} w_j a_{ij}\right)^2 \tag{3}$$

where y_i is the actual bid for the i^{th} auction
m is the total number of attributes
w_j is the regression coefficient for the j^{th} attribute
a_{ij} is the j^{th} attribute for the i^{th} auction.

2.3.2. *Non-parametric approach to Price Prediction*

A *k*-Nearest Neighbour approach is used to identify *k* auctions in the dataset that are similar to the target auction whose price is to be predicted. This is a simple classification method that does not make assumptions about the form of the relationship between the closing price of the auction and the predictor variables. This is a non-parametric method because it does not involve estimation of parameters (coefficients) as a linear regression technique would; instead, this method draws information from similarities between the predictor values of the different auctions in the dataset based on the distance between these auction records.

The *k-NN* is employed based on the distinct clusters' characteristics. Euclidean distance (d_{ij}) is used to measure the distance between the target auction and the auctions in the respective cluster (the one to which it belongs). The auction data is normalised before computing the distance to ensure that the distance measure is not dominated by variables with a large scale.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$d_{ij} = \sqrt{\left(OpenB_i - OpenB_j\right)^2 + \left(CloseP_i - CloseP_j\right)^2 + \left(NUM_i - NUM_j\right)^2 + \left(AvgB_i - AvgB_j\right)^2 + \left(AvgBR_i - AvgBR_j\right)^2} \quad (4)$$

where d_{ij} is the Euclidean distance between the i^{th} and j^{th} auction.

After computing the distance between the auctions, the target auction is assigned to a class of k auctions based on the classes of its auction neighbours. A value of k is chosen which achieves a balance between over-fitting to the predictor information (if k is too low) and ignoring this information completely (if k is too high). To find the optimal choice of k, the auctions in the training dataset are used to classify the auctions in the validation dataset and the error rates for each value of k are calculated. A value of k is chosen which has the best classification performance and so minimises the misclassification rate (of the validation set).

The closing price of the target auction is predicted using the weighted average of the closing price of the k nearest neighbour.

$$CloseP_{TA} = \frac{\sum_{i=1}^{k} w_i CloseP_i}{\sum_{i=1}^{k} w_i} \quad (5)$$

where k auctions are ordered as per their increasing distances from the selected auction, w_i is the weight assigned to the i^{th} auction and $w_1 > w_2 > w_3 > \ldots\ldots w_k$, $CloseP_{TA}$ is the closing price of the target auction.

3. Method Validation

The price prediction methodology selects an auction to participate in and assesses the value of the auctioned item. An auction for participation is selected by partitioning similar auctions using a clustering based approach and the value of an item is assessed using parametric and non-parametric machine learning techniques. The methods for clustering auctions and value assessment are validated separately for accurate end price prediction results.

3.1. *Validation of the Auction Clusters*

The resulting clusters are validated using two criteria: cluster separation and cluster interpretability. *Dunn's Index (DI)* is applied as a measure to identify "compact and well separated" (*CWS*) clusters.(Halkidi, Batistakis, & Vazirgiannis, 2001; Rivera-Borroto, Rabassa-Gutiérrez, Grau-Abalo, Marrero-Ponce, & García-de la Vega, 2012)
Dunn's index (DI) is defined for k clusters as below:

$$DI = \min_{1 \le m \le k} \left\{ \min_{1 \le n \le k, m \ne n} \left\{ \frac{d(m,n)}{\max_{1 \le c \le k} d'(c)} \right\} \right\} \quad (6)$$

where $d(m,n)$ represents the distance between clusters m and n and is defined as

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$d(m,n) = \min_{x \in m,\, y \in n} d(x,y) \tag{7}$$

$d'(c)$ measures the intra-cluster distance of cluster *c* and is defined as

$$d'(c) = \max_{x,\, y \in c} d(x,y) \tag{8}$$

For compact and well-separated clusters in the dataset, the distance between the clusters is expected to be greater and the intra-cluster distance is expected to be small. *DI > 1* shows that the dataset is being partitioned into compact and well- separated clusters(Rivera-Borroto et al., 2012).

Further, for the interpretation of each cluster: a) its characteristics are explored by obtaining summary statistics from each cluster on each attribute used in the cluster analysis and b) it is categorised based on their characteristics.

3.2. *Validation of the Value Assessment*

The value of the item is assessed by predicting the closing price of its auction using parametric and non-parametric approaches based on machine learning algorithms. Multiple linear regression is used as the parametric approach, and *k*-nearest neighbour is used as the non-parametric approach. These methods are validated separately for accurate results.

The multiple linear regression method for price prediction is validated by dividing the data into two distinct sets: a training dataset and a validation dataset. These sets represent the relationship between the dependent and independent variables. The training dataset is used to estimate the regression coefficients, and the validation dataset is used to validate these estimations. The prediction is made for each case in the validation data using the estimated regression coefficients and then these predictions are compared to the value of the actual dependent variable to validate the outcomes. The square root of the mean of the squares of these errors is used to compare different models and to assess the prediction accuracy of the method used.

To validate the *k*-nearest neighbour method, first, the auctions' dataset is divided into training and validation datasets and then *k*-closest members (based on the minimum distance between them) of the training dataset are located for each auction in the validation dataset. *k* models are built and scoring on the best of these models is performed to choose the value of *k*. The value of *k* is chosen which has the best classification performance while classifying the auctions in the validation data using the auctions in training data. A very low value of *k* classifies data sensitive to the local characteristics of the training data, and a large value of *k* predicts the most frequent recorded type in the dataset. To find the optimal *k*, the mis-classification rate of the validation data is examined for different values of *k* and the one is chosen which minimises the classification rate. The *k*-nearest neighbour method is validated by comparing the square root of the average of the squared residuals of different models to assess the prediction accuracy.

4. Experiments and Observations

Several experiments were performed using the end price prediction methodology using eBay auctions dataset in order to demonstrate the validity of the proposed approach. It is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

important to note that the main purpose of these experiments was to select an auction where participation would give maximum surplus and to assess the true value of the auctioned item. An auction was selected using a clustering-based methodology with a bid mapping and selection approach, and the value of the item was assessed using machine learning techniques. The dataset used and the experiments for auction selection and value assessment are presented below.

4.1. *Dataset Used*

The dataset includes the complete bidding records for 149 auctions of new Palm M515 PDAs. This dataset was made available at httt://www.ModelingOnlineAuctions.com by researchers in 2010.(Wolfgang Jank & Shmueli, 2010) All the auctions in this dataset are for the same product: a new Palm M515 handheld device. The market price of the item at that time was $250. The bidding record of each auction includes the auction ID, opening bid amount, the closing price of the auction, information about the ratings of bidders and the seller, the number of bids placed in the auction, all bids along with their placement and the duration of the auction. Table 1 shows statistical information for the dataset. The opening bid is set by the seller, influencing the number of bidders attracted to the auction. As the number of bidders in the auction rises, it becomes more competitive, and the closing price of the item is also higher. The average number of bids is 21.36 and the average bid amount of the auction is $148.91.

Table 1. Description of data

Variable	***Min***	***Max***	***Mean***	***Std. Deviation***
Opening Bid(OpenB)	0.01	240	40.55	69.71
Closing Price(CloseP)	177	280.65	228.24	16.10
# Bids(NUM)	2	51	21.36	10.16
Avg bid amt(AvgB)	77.47	243.75	148.91	33.13
Avg bid rate(AvgBR)	-2196.27	42679.94	11624.09	8143.42
OpenB: Opening Bid or the starting price of an auction. *CloseP: Closing Price or the end price of an auction.* *NUM: Total number of bids placed in an auction.* *AvgB: Average Bid amount of each auction.* *AvgBR: Average Bid Rate of the auction*				

4.2. *Experiments for Auction Selection*

Auctions of the same or similar items were clustered using the *k-means* algorithm. The value of *k* in *k*-means algorithm was estimated by the Elbow approach using one-way analysis of variance (ANOVA). The percent of variance was calculated after performing clustering for subsequent values of *k* using the *k*-means algorithm to estimate the point where marginal gain drops. In the experiments, this point occurred after five clusters, as shown in Figure 2. The input space was, thus, divided into five clusters, considering a set of attributes of auctions comprising the opening bid, closing price, number of bids, average bid amount and average bid rate for a particular auction. These five clusters respectively contained 19%, 34%, 20%, 21% and 6% of the auctions' data (Table 2). Based on the transformed data after clustering and the characteristics of the current

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

auctions, the BMS algorithm nominated the cluster for each of the ongoing auction to select the auction in which to participate. The algorithm used *AvgBR* value at the beginning of the last hour to map ongoing auctions to the clusters for auction selection.

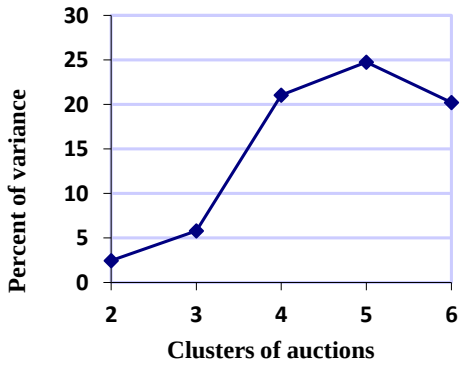


Fig. 2. Choosing the value of *k* for auction input data

Table 2. Attributes' statistics for each cluster

Cluster no.	*%age of Auctions' data*	*Normalized AvgBR*	*Avg. NUM*	*Avg. OpenB*	*Avg. CloseP*
1	19%	8.30	11.07	112.14	225.02
2	34%	21.69	22.53	28.31	226.65
3	20%	35.17	27.86	16.09	231.73
4	21%	53.20	21.96	24.91	225.83
5	6%	86.48	23.12	20.37	227.80

The clusters obtained using cluster analysis, were validated using two criteria: cluster separation and cluster interpretability, as explained in Section 3.1. *Dunn's Index (DI)* was calculated to find *CWS* clusters. In these experiments, *DI > 1* showed that the dataset was being partitioned into the *CWS* clusters. The summary statistics from each cluster for each attribute are given in Table 2. It is obvious that the *k-means* clustering algorithm

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

distinguished these five clusters according to the *AvgBR* of the auctions. We interpret these clusters with *very low* (8.30), *low* (21.69), *medium* (35.17), *high* (53.20) and *very high* (86.48) average bid rate (AvgBR) auctions. These clusters were ordered according to the increasing values of their *AvgBR* from *very low* to *very high* for the sake of convenient representation of results. It is also evident from Table 2 that the auctions with the highest opening bid attracted the least bidders (cluster1). In addition, the bid rate of these auctions was lowest, which lowered the closing price of the auction. On the other hand, the lowest value opening bid attracted the most bidders, which in turn raised the closing price of these auctions even with a medium bid rate (cluster 3).

It is noteworthy that in 92 of the 149 auctions in this dataset, the winner first appeared in the last hour of the auction; this accounted for 62% of total auctions, a finding consistent with the recognition of the late-bidding attitude of bidders in the online auction literature.(Du, Chen, & Bian, 2010; Ockenfels & Roth, 2006; Rasmusen, 2006) As explained, clustering divided the auction data into groups of auctions with a distinct range of average bid rate (*AvgBR*) values. It was observed that the value of *AvgBR* in 78% of the completed auctions belonged to the same cluster as at the beginning of the last hour of the auction. These criteria serve as benchmarks in validating the procedure adopted by the *BMS* algorithm which maps ongoing auctions to the clusters based on their *AvgBR* value at the beginning of the last hour.

4.3. *Experiments for Value Assessment*

The predictive accuracy of the models for value assessment of the auctioned item was compared using *RMSE* (root mean square error) and *MRE* (mean relative error) error measures. For this purpose, *RMSE* was the square root of the average squared errors and *MRE* was defined as how much the prediction departed from the real price on average. The experiment was repeated four times to derive stable evaluation results, and in each of these subsequent experiments, the dataset was randomly split into training and validation sets. The average over these sets is reported as the prediction results. The *RMSE* for each cluster and average *MRE* for all the clusters are reported for evaluation in Table 3 and Table 4 respectively. The results show that the parametric approach to price prediction is not as promising as the non-parametric approach since the non-parametric approach performs better when the predicted variable can be defined by multiple combinations of predictor values allowing for a wide range of relationships between the predictors and the response. Further, in contrast to parametric linear regression functions, the non-parametric approach is able to determine which of the attributes should be used for modelling the relationship with the target variable. It was observed that the *k*-nearest neighbour technique performed better, yielding minimum *RMSE* and *MRE* on average, so this technique was used for value assessment.

The price predicted for valuation purposes was validated by comparing the *RMSE* in two scenarios: first, the price was predicted when considering the input auctions as a whole, and second, it was predicted using different auction clusters. The results were evaluated by comparing the root mean square errors (*RMSEs*) in both of these scenarios. The results of the experiment demonstrated an improvement in RMSE by 36.64% on average when the price is predicted using different auction clusters.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 3. Root mean square error for price prediction approaches

	Parametric approach	***Non-parametric approach***
Cluster1	15.15	12.76
Cluster2	14.42	13.82
Cluster3	15.65	6.58
Cluster4	14.98	13.84
Cluster5	54.29	4.69

Table 4. Average mean relative error for price prediction approaches

	Parametric approach	***Non-parametric approach***
Average MRE	0.09	0.04

4.4. *Comparison with Other Algorithms*

The previous section described experiments performed on the Palm PDA dataset of eBay auctions for auction selection and value assessment. The value of an item was assessed by predicting the closing price of its auction. Other studies have also used non-parametric techniques to predict the closing price of eBay auctions and their findings can be compared with the clustering-based non-parametric approach to closing price prediction in this paper.

D. Van Heijst et al. validated the price prediction technique using heterogeneous and homogeneous datasets.(Van Heijst et al., 2008) Their study used datasets from closed Nike and Canon auctions as heterogeneous datasets. The Nike dataset included data for auctions of both used and new models of Nike men's shoes, while the Canon dataset contained data about various camera models as well as accessories like lenses and batteries. Datasets for auctions of H700 Motorola Bluetooth headsets and 30 GB Apple iPod mp3 players were included as homogeneous datasets. In these datasets, the items were technically identical, a trait also true for the dataset for auctions of new Palm M515 PDAs described above. We can, thus, turn to Van Heijst et al's closing price prediction results (using the dataset of 30 GB Apple iPod mp3 players) as a comparison for this study's clustering-based non-parametric approach to closing price prediction. The proposed price prediction methodology is also compared with the price prediction results of I. Raykhel and D. Ventura, who used laptop auctions as a homogeneous dataset.(Ilya Raykhel & Ventura, 2009)

The results achieved by these other researchers are compared with the clustering-based price prediction approaches in Table 5. It can be seen that the clustering-based approach is the most effective in predicting the closing price of auctions. The *MRE* recorded using the clustering-based approach (0.04) is lower than the results achieved with the algorithms of D. Van Heijst et al. (0.1) and I. Raykhel and D. Ventura (0.164),

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

suggesting the greater precision of the proposed methodology.(Ilya Raykhel & Ventura, 2009; Van Heijst et al., 2008)

Table 5. Error comparison using others' algorithms

	Clustering-based price prediction	***I. Raykhel and D. Ventura(Ilya Raykhel & Ventura, 2009)***	***D. Van Heijst et al.(Van Heijst et al., 2008)***
MRE	0.04	0.164	0.1

5. Conclusions and Future Work

In this paper we presented a design of end price prediction methodology which chooses a target auction to compete in from simultaneous auctions for the same or similar items and assesses the value of the item on offer. To achieve these aims, a clustering approach is adopted along with a bid mapping and selection technique.

Price dynamics—the path taken by bid amounts over the course of an auction—is carefully considered in this study. This is one of the main contributions to decision-making when it comes to estimating an auction's closing price that is used, in turn, to assess the value of the auctioned item. It is unrealistic to assume that the price dynamics of simultaneous auctions for the same or similar items remain the same across any auction environment. This study has therefore characterised auctions of the same or similar items based on their price dynamics before selecting an auction to compete in and assessing the true value of the auctioned item. A bid mapper and selector (BMS) algorithm has been presented which chooses a target auction to compete in. Machine learning techniques are used to estimate the closing price.

This closing price prediction method for value assessment has been validated using a dataset from eBay auctions for a new Palm M515 PDA. Overall, the presented methodology has produced a prediction model superior in accuracy to the closing price prediction methodologies of I. Raykhel and D. Ventura and D. Van Heijst et al. (Ilya Raykhel & Ventura, 2009; Van Heijst et al., 2008)

For closing price prediction, this study has used a *K-NN* algorithm, which could be improved further by accelerating the process to find the nearest neighbour for a large training dataset. In future, two approaches are planned to speed up the nearest neighbour classification step: first, sophisticated data structures such as search trees will be applied since these take an "almost nearest neighbour" approach to classification, and second, redundant points will be edited out of the training data as these have no effect on the classification and are surrounded by records that belong to the same class.

This study for closing price prediction has focused on auctions with hard closing rules only. It would also be interesting to explore the possibility of adapting the clustering based bid mapping and selection technique to auctions with soft closing rules and comparing the performance with hard-closing-rules auctions for the same item. This evaluation could be done using a paired design where identical items are auctioned at the same time, with one item in the pair sold off in a hard-closing-rules auction and the other in a soft-closing-rules setting.

References

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Bajari, P., & Hortacsu, A. (2003). The winner's curse, reserve prices, and endogenous entry: empirical insights from eBay auctions. *RAND Journal of Economics*, 329-355.

Bapna, R., Goes, P., & Gupta, A. (2003). Analysis and design of business-to-consumer online auctions. *Management Science, 49*(1), 85-101.

Byde, A., Preist, C., & Jennings, N. (2002). *Decision procedures for multiple auctions.* Paper presented at the Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2.

Dass, M., Jank, W., & Shmueli, G. (2011). Maximizing bidder surplus in simultaneous online art auctions via dynamic forecasting. *International Journal of Forecasting, 27*(4), 1259-1270. doi: http://dx.doi.org/10.1016/j.ijforecast.2011.01.003

Du, L., Chen, Q., & Bian, N. (2010). *An empirical analysis of bidding behavior in simultaneous ascending-bid auctions.* Paper presented at the International Conference on E-Business and E-Government (ICEE), .

Ghani, R., & Simmons, H. (2004). *Predicting the end-price of online auctions.* Paper presented at the Proceedings of the International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management, Pisa, Italy.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2-3), 107-145.

Haruvy, E., & Popkowski Leszczyc, P. T. L. (2010). Internet Auctions. *Foundations and Trends® in Marketing, 4*(1), 1-75.

Jank, W., & Shmueli, G. (2005). Profiling price dynamics in online auctions using curve clustering (pp. 1-28): Smith School of Business, University of Maryland.

Jank, W., & Shmueli, G. (2010). *Modeling Online Auctions* (Vol. 91): Wiley. com.

Jank, W., & Zhang, S. (2011). An automated and data-driven bidding strategy for online auctions. *INFORMS Journal on computing, 23*(2), 238-253.

Kaur, P., Goyal, M., & Lu, J. (2012). An Integrated Model for a Price Forecasting Agent in Online Auctions. *Journal of Internet Commerce, 11*(3), 208-225.

Kehagias, D., Symeonidis, A., & Mitkas, P. (2005). Designing pricing mechanisms for autonomous agents based on bid-forecasting. *Electronic Markets, 15*(1), 53-62.

Ku, G. (2000). Auctions and auction fever: Explanations from competitive arousal and framing. *Kellogg Journal of Organization Behavior, 2000*, 1-41.

Ockenfels, A., Reiley, D. H., & Sadrieh, A. (2006). Online auctions *Economics and Information Systems Handbook* (pp. 571-622). Amsterdam: Elsevier Science.

Ockenfels, A., & Roth, A. E. (2006). Late and multiple bidding in second price Internet auctions: Theory and evidence concerning different rules for ending an auction. *Games and Economic Behavior, 55*(2), 297-320.

Park, Y. H., & Bradlow, E. T. (2005). An integrated model for bidding behavior in Internet auctions: Whether, who, when, and how much. *Journal of Marketing Research, 42*(4), 470-482.

Rasmusen, E. B. (2006). Strategic implications of uncertainty over one's own private value in auctions. *BE Press Journal, 6*(1), Article 7.

Raykhel, I., & Ventura, D. (2008). *Real-time automatic price prediction for ebay online trading*: Brigham Young University. Dept. of Computer Science.

Raykhel, I., & Ventura, D. (2009). *Real-time automatic price prediction for ebay online trading.* Paper presented at the Proceedings of the Innovative Applications of Artificial Intelligence Conference.

Rivera-Borroto, O. M., Rabassa-Gutiérrez, M., Grau-Abalo, R. d. C., Marrero-Ponce, Y., & García-de la Vega, J. M. (2012). Dunn's index for cluster tendency assessment of pharmacological data sets. *Canadian journal of physiology and pharmacology, 90*(4), 425-433.

Roth, A. E., & Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the Internet. *The American Economic Review, 92*(4), 1093-1103.

Sun, E. (2005). The effects of auctions parameters on price dispersion and bidder entry on eBay: a conditional logit analysis: working paper, Stanford University.

Van Heijst, D., Potharst, R., & Van Wezel, M. (2008). A support system for predicting eBay end prices. *Decision Support Systems, 44*(4), 970-982.

Wang, S., Jank, W., & Shmueli, G. (2008). Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economic Statistics, 26*(2), 144-160.

Wilcox, R. T. (2000). Experts and amateurs: The role of experience in Internet auctions. *Marketing Letters, 11*(4), 363-374.

Xuefeng, L., Lu, L., Lihua, W., & Zhao, Z. (2006). Predicting the final prices of online auction items. *Expert Systems with Applications, 31*(3), 542-550.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Thinging Ethics for Software Engineers

Sabah S. Al-Fedaghi
Computer Engineering Department
Kuwait University
City, Kuwait
sabah.alfedaghi@ku.edu.kw

***Abstract*—Ethical systems are usually described as principles for distinguishing right from wrong and forming beliefs about proper conduct. Ethical topics are complex, with excessively verbose accounts of mental models and intensely ingrained philosophical assumptions. From practical experience, in teaching ethics for software engineering students, an explanation of ethics alone often cannot provide insights of behavior and thought for students. Additionally, it seems that there has been no exploration into the development of a conceptual presentation of ethics that appeals to computer engineers. This is particularly clear in the area of software engineering, which focuses on software and associated tools such as algorithms, diagramming, documentation, modeling and design as applied to various types of data and conceptual artifacts. It seems that software engineers look at ethical materials as a collection of ideas and notions that lack systemization and uniformity. Accordingly, this paper explores a thinging schematization for ethical theories that can serve a role similar to that of modeling languages (e.g., UML). In this approach, thinging means actualization (existence, presence, being) of things and mechanisms that define a boundary around some region of ethically related reality, separating it from everything else. The resultant diagrammatic representation then developed to model the process of making ethical decisions in that region.**

Keywords-ethics; software engineering; conceptual modeling; ethical theory; diagrammatic representation; thinging

I. INTRODUCTION

The increasing reliance on computers for infrastructure in modern society has given rise to a host of ethical, social, and legal issues such as those of privacy, intellectual property, and intellectual freedom. These issues have increased and became more complex. Making sound ethical decisions is thus an important subject in computer and software engineering [1]. Computer professional societies (e.g., ACM and IEEE-CS) have proposed several codes of ethics, including the Code of Ethics and Professional Practice for software engineers [2-3] as the standard for teaching and practicing software engineering [4]. The Association for Library and Information Science Education [5] specifies that student learning outcomes in studying ethics include (1) recognizing ethical conflicts; (2) developing responsibility for the consequences of individual and collective actions; and (3) ethical reflection, critical thinking, and the ability to use ethics in professional life.

The field of computer ethics is changing rapidly as concerns over the impact of information and communication technology on society mount. New problems of ethics emerge one after another, and old problems show themselves in new forms. Ethics is often tied to legal procedures and policies that, if breached, can put an organization in the midst of trouble [6].

In this context, we adopt the classical engineering method of *explaining* a phenomenon through modeling. For example, in developing system requirements, a model-based approach is used to depict a system graphically at various levels of granularity and complexity. The resultant unified, conceptual model facilitates communication among different stakeholders such as managers, engineers, and contractors and establishes a uniform vocabulary that leads to common understanding and mental pictures of different states of the system. Similarly, in teaching *modeling* and *explaining* are two closely related practices and "multiple models and representations of concepts" [7] are used to show students how to solve a problem or interpret a text. Representations and models are used in building student understanding [7].

This paper proposes diagrammatically modeling decision making in ethical systems based on the framework of *thinging* wherein things *thing* (a verb that refers to "manifest themselves in the system of concern"), then complete their life cycles though processing, transferring and receiving. The model includes the things and their machines that create, process, release, transfer and receive things.

Ethical systems are usually described as principles for distinguishing right from wrong and beliefs about proper conduct. Ethical topics are complex, with excessively verbose accounts of mental models and intensely ingrained philosophical assumptions. For example, some studies have found that many IT students are unable to distinguish criminal actions from unethical behavior [8].

From practical experience of teaching computer ethics to computer/software engineers (text is Johnson's "Computer Ethics" [9]), it is observed that a textual explanation of ethics often cannot provide insights of ethical behavior and thought for students. Additionally, it seems that there has been no exploration into the style of an ethical *conceptual model* that appeals to computer and software engineers. A conceptual model is an abstraction that describes things of interest to systems. It provides an exploratory basis for understanding and explanation of the phenomenon under consideration. The model can be used as a common representation to focus communication and dialogue, especially in pedagogic environments.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Specifically, we aim to target software engineering students, as software professionals have the power to do good or bad for society and we need to use their knowledge and skills for the benefit of the society [10]. According to the IEEE Computer Society and Association for Computing Machinery (ACM) code of ethics [2-3], every software professional has *obligations* to society, self, profession, product, and employer. Software engineering involves such topics as computer programs and their associated tools such as algorithms, diagramming, documentation, modeling, and designing as applied to various types of data and conceptual artifacts. From other directions and as observed in actual experience of teaching ethics, it seems that software engineers look at ethics as an "alien" topic that contains a collection of ideas and notions that lack systemization and uniformity. Systemization here refers to systems (a highly regarded engineering notion) and their notations, as in representing a system in terms of the classical input-process-output (IPO) model.

Accordingly, this paper explores a thinging-based schematization for ethical theories by expressing ethics in a familiar style for software engineers that is similar to modeling languages (e.g., UML [11]). Since the ability to make diagrams is a valuable and a common skill for programmers skill in software engineering, the proposed method aims at improving students' abilities to describe principles of ethics, to apply a model for ethical decision-making, and to practice diagrammatic communication activities.

Note that this is not a paper in the field of ethics; rather, it introduces a diagrammatic language to describe ethical notions. Consequently, the ethics theories that will be given, if they include errors from the point of ethics, can be corrected by modifying the diagrams without affecting the aim of the paper.

In the next section, we will briefly explain our main tool of modeling, called *Thing Machines* (TM) [12-21]. The example in the section is a new contribution. In Section 3, a brief description of ethics theories is introduced. Sections 4 and 5 give two sample applications of TM. In Section 4, we apply TM to the ethical system of Kantism. We show that TM representation provides a new method of utilizing diagrams to analyze ethics. Motivated by the teaching environment at Kuwait University, we also apply TM to Islamic ethics.

II. Thinging Machines

In philosophy, Thinging refers to "defining a boundary around some portion of reality, separating it from everything else, and then labeling that portion of reality with a name" [22]. According to Heidegger [23], to understand the thingness of things, one needs to reflect on how thinging expresses how a "thing things" that is "gathering", uniting, or tying together its constituents, just as the bridge makes the environment (banks, stream, and landscape) into a unified whole. From slightly different perspectives, thinging and things *thing* (verb) refer to *wujood*: actualization (manifestation), existence, being known or recognized, possession of being, being present, being there, entity (a creature), appearance, opposite of nothingness. For example, a number (in abstract) *has wujood*, but it has no *existence*.

In our approach, there is a strong association between systems and their models. A system is defined through a model. Accordingly, we view an ethical system as a system of "things of ethics." The system also *things* itself by machines of these things. In simple words, as will be exemplified later, it is a web of (abstract) machines represented as a diagram (the grand machine). A machine can thing (create), process, receive, transfer, and/or release other things (see Fig. 1). These "operations" are represented within an abstract Thinging Machine (TM) as shown in Fig. 1.

A. Example

According to Rosnay [24], the most complete definition of a system is that it is a set of elements in dynamic interaction organized for a goal. Rosnay [24] presents a diagrammatic representation of a reservoir system that fills and empties water that is maintained at the same level as shown in Fig. 2. The figure contains the basic notion that can be used in building the so-called the Thinging Machine (TM) model. Fig. 3 illustrates the notions of things and machines in the reservoir system that will be used in the TM model.

Fig. 4 shows such a system using TM. The water as a thing flows from the outside (circle 1 in the figure) through the valve to the reservoir (2) and outside (3).

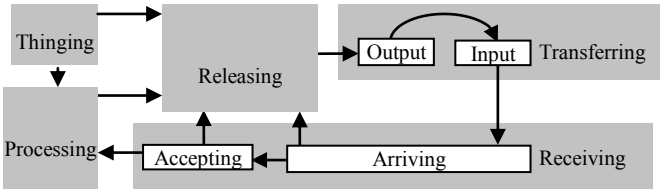


Fig. 1. Thinging machine.

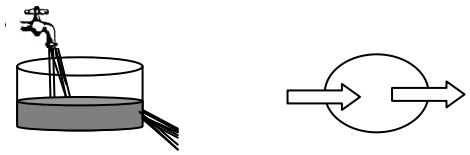


Fig. 2. Diagrammatic representation of a water reservoir system (Redrawn, partial from Rosnay [24])

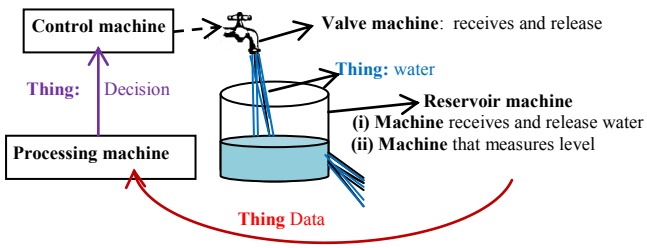


Fig. 3. A water reservoir system and its things/machines.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

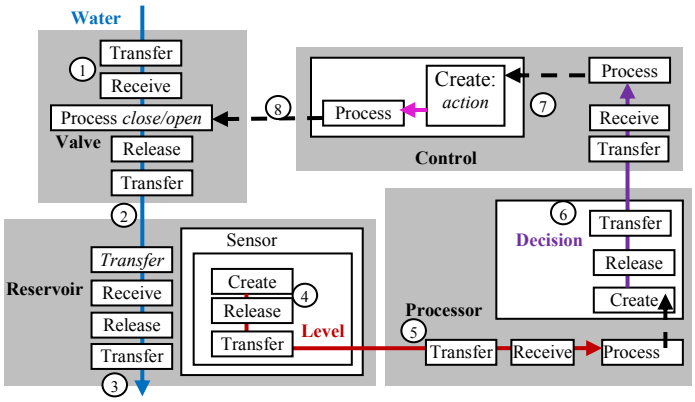


Fig. 4. The TM model of the water reservoir system.

The reservoir machine includes a sensor sub-machine that measures the water level (4). The measurement data flow to a processor (5) that triggers a decision machine (the dashed arrow) to generate a control decision (6). The decision flows to control machine of the valve (7), which triggers opening or closing the flow of water from the outside source (8).

An *event* in the TM model is a machine that includes the event itself (sub-machine) and the content of the event, which includes the time sub-machine and region sub-machine of the event at a minimum. Fig. 5 shows the representation of the event *The level measurement is generated and sent to the processor*. Note that the *region* is a sub-graph of the TM static representation of the water reservoir system. For simplicity's sake, we will represent events by only their regions.

Accordingly, Fig. 6 shows meaningful events in the static representation of the reservoir system. Fig. 7 shows the chronology of events that can be used to build a control for the system.

B. Thinging

We can say that a thing is a *machine* that things (verb), including creating other things that, in turn, are machines that produce things as illustrated in Fig. 8. The chicken is a thing that flows out of the egg. It is also a machine that things (creates), processes (changes), receives, releases, and/or transfers things. In Fig. 8, the chicken machine things (creates) eggs in addition to other things not shown in the figure (e.g., cluck machine, waste machine).

Going by the function of a TM, we define a thing as follows:

A thing is what manifests itself in the creation, processing, receiving, releasing, and transferring stages of a thinging machine.

Accordingly, in a TM, we have five kinds of thinging: the machine *creates* (in the sense of *wujood* explained above), *processes (changes)*, *arrives*, *transfers*, and releases (things *wait for departure*). Thinging is the emergence, changing, arriving, departing, and transferring of things.

The utilization of thinging in this paper is not about the philosophical issues related to the ontology of things and their nature; rather, it concerns the representation of things in and machines in a system.

The TM's definition of "thing" broadens its characterization (in comparison to the ontological base of Heidegger [23]'s thinging) by including other thinging aspects: process-ness, receive-ness, transfer-ness, and release-ness. All four features form possible "thingy qualities" [22] after *wujood* (the appearance) of the thing in the grand machine (system of concern).

A thing that has been created refers to a thing that has been born, is acknowledged, exists, appears, or emerges as a separate item in reality or system and with respect to other things. A factory can be a thing that is constructed and inspected as well as a machine that receives other things (e.g., materials) to create products. A factory is a thing when it is processed (e.g., created), and it is a machine when it is processing things (e.g., creating products).

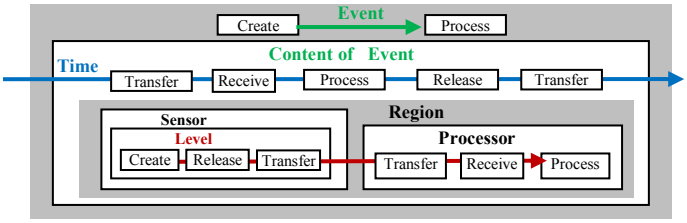


Fig. 5. The event *The level measurement is generated and sent to the processor*.

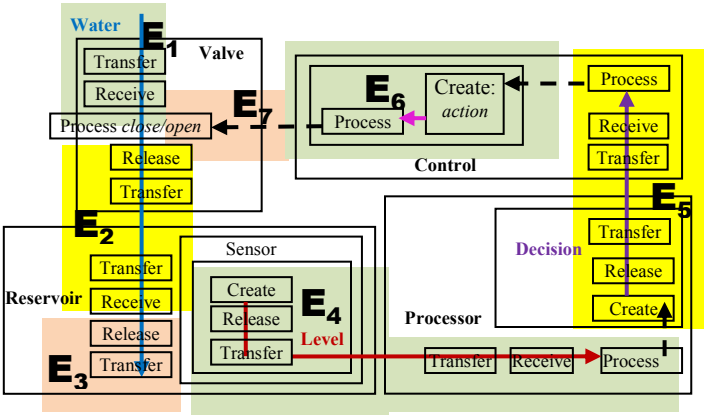


Fig. 6. The TM event-ized representation of the water reservoir

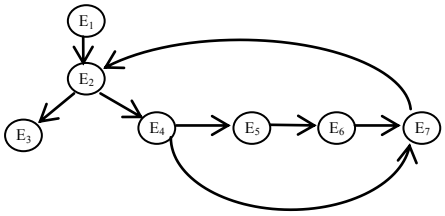


Fig. 7. The chronology of events in the reservoir system.

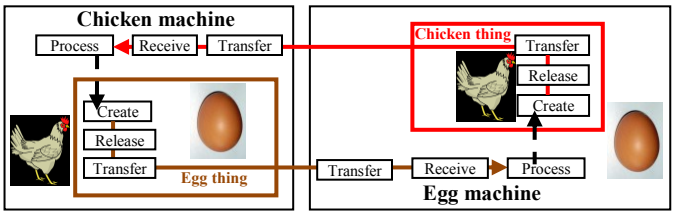


Fig. 8. Illustration of a machine that can be a machine and a machine that can be a thing.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

To *Create* a thing means that it *comes about*, and this implies the possibility of its un-thinging (the opposite of *wujood*) within a machine. A collection of machines within a thing forms a larger machine. The stomach is a food-processing machine in the larger digestive machine. The digestive system is one machine in the human being machine whose function is related to the thing 'food.' which is digested (processed) to create waste. A human being is a thing in a school machine.

Processing refers to a change that a machine performs on a thing without turning it into a new thing (e.g., a car is processed when its color is changed).

Receiving is the flow of a thing to a machine from an outside machine. *Releasing* is exporting a thing outside the machine. It remains within the system, labeled as a released thing, if no release channel is available. *Transferring* is the released thing departing to outside the machine.

The world of a TM consists of an arrangement of machines, wherein each thing has its own unique stream of flow. TM modeling puts together all of the things/machines required to assemble a system (a grand machine).

III. Ethical Theories

Theories provide explanations to laws. In science, laws manifest regularities in natural phenomena. Scientific laws can be discovered using reason as it is applied to experience. *Natural* laws tell *what is thinging* or *will be thinging*. *Moral* laws are related to *what thinging ought to be*. Instead of natural phenomena, they manifest regularities *in human thinging* (manifestation of oneself or behavior). If human beings were wholly rational beings, moral thinging would be like natural thinging. But since we have inclinations and desires, moral laws appear as imperatives. Nevertheless, there is the claim that morality can be based on natural law. This refers to a system of law that is intrinsic to the structure of the universe [25].

"Being related to ethics" typically refers to thing decisions according to some ethics principles. Normative ethical theories are used to thing ethics judgment when deciding among several alternative courses of response. Ethics determination involves decision-making. Is decision-making in ethics different from other kinds of decision-making, such as in law, rule, policy-making, etc.?

It is often said that moral situations are usually vague regarding which principles are applicable to them. Moral principles may conflict with each other, creating moral dilemmas. Disagreements may also arise about how to interpret and apply these principles in particular situations [26-27]. However, this characteristic of moral situations does not mean that the ethical decision process is fundamentally different from any other decision process that involves vagueness and uncertainly. The decision-making process comprises input, process, and output (IPO mode) and includes factors affecting the determination.

When an ethics problem presents itself, an evaluation machine involves several things, such as objects, persons, circumstances, events, and acts. Ethics values resulting from such a process depend on these factors [28-30].

Actions are morally right when they comply with a moral principle or duty. Thus, in general we have an intended act and an ethical machine that give an act value to an agent who decides which value he/she chooses.

IV. Modeling Kantism

Ethical theories are divided according to the nature of moral standards used to decide whether a given conduct is right or wrong. Two main categories of normative theories can be identified: the teleological (consequentialist) theories and deontological theories. "Telos" and "deon" in Greek mean "end" and "that which is obligatory", respectively [31]. Deontology is based on the primacy of duty over consequences, where some of which are morally obligatory. Obligation is not necessarily a deontological characteristic. A utilitarian theory, for example, may utilize the concept of obligation teleologically (see [32]). Actions are morally right when they comply with a moral principle or duty. In this paper, we exemplify the application of TM to ethical decision making to two systems, Kantism and Isla.

A. Applying TM to Kantism

In Kantism, moral obligations must be carried without qualification, and these must hold for everyone without exception. Hence, the form that moral principles must take is law-like, which can provide the basis for morality. See [28], [33-34]. Here, the will is the human capability to make a decision based on reason. Thus, we should act according to rules that we can embrace as universal laws. Moral principles are categorically (without regard to consequences or exceptions) binding. Humans as rational beings are also moral beings who understand what it takes to live as such. Hence, we impose morality on ourselves, and no one else, as in the case where no one can force us to be rational. As we freely choose to be rational and accept rationality, we also freely choose to be moral and accept morality.

The basic principle categorical imperative (CI) is a modification of the Golden Rule [35] as follows:

Take an action if its maxim (general principle of conduct) were to become a universal law through your will; i.e., you want others to treat it as a moral law.

Maxims, according to Kant, are subjective rules that guide action. An ethical decision is universal, applied consistently across time, cultures, and societal norms.

According to Pascal [36], Kantian ethics have great influence on many thinkers, such as Habermas [37], who proposed that action should be based on communication and Rawls' social contract theory [38]. Pascal [36] modelled the "categorical imperative" as shown in Fig. 9. McKnight [39] in her "The Categorical Imperative for Dummies" uses another type of diagram, as shown in Fig. 10.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

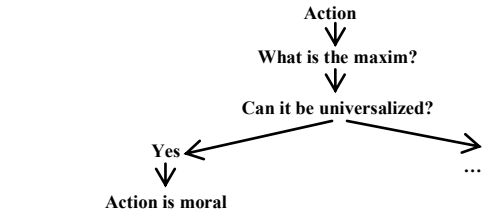


Fig. 9. Categorical Imperative model (Re-drawn, partial from Pascal [36]).

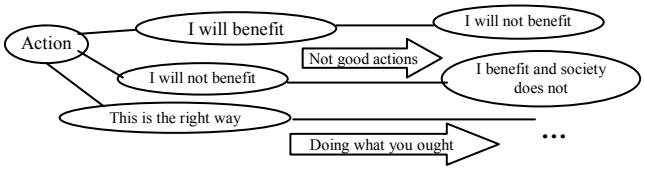


Fig. 10. Kant's theory of duty (re-drawn, partial from McKnight [39]).

According to the Moral Robots blog [40], an action is morally right if it has a good motivation and conforms to Kant's categorical imperative, as explained in Fig. 11. UML use case diagrams [11] have also been used in presenting a method to decide to grant patients' requests for access to their health information based on Health Information Privacy Code, and national and international codes of ethics [41].

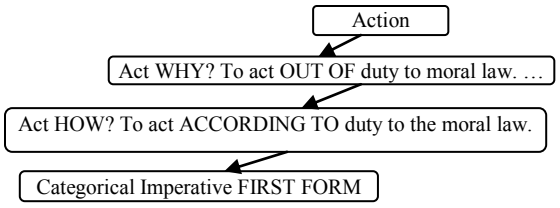


Fig. 11. Kant's Ethics (re-drawn, partial from Moral Robots [40]).

These types of diagrams point to the need for more systematic diagrammatic methods to facilitate explaining ethics. *Systematic* here refers to following a defined methodological approach and an explicitly defined process [42].

Fig. 12 shows the TM representation of Kant's categorical imperative: *Take an action if its maxim were to become a universal law through your will*, and Fig. 13 shows the corresponding dynamic model of such a method of judging actions ethically.

In Fig 12, a person things (i.e., creates) an intended action in his/her mind (circle 1), and this triggers thinging a universe (2) that includes him/herself and others (3—at least two other persons). Each person, I, other-1 and other-2 creates the intended action that flows to the other two in the universe. Processing such a universe (4) triggers the *will* to be in the state of agreeing/disagreeing with such a universe. If the person wants others to treat the intended action as a moral law, then he/she would process the action (5 - determine how to realize the action) and then implement it (6). The copy model (7) guarantees that the mental universe is feasible in reality.

Events in Fig. 13 are created based on meaningfulness to the modeler. Note the time and space machines at the bottom of the figure. The time of the mental universe is received and processed (takes its course) but it never ends (no release and transfer in such a universe). The execution of the action goes on all the time among I, other-1 and other-2.

Additionally, this happens everywhere. It may be interesting to consider that space itself flows as a thing through the repeated events. Suppose that an airplane flows from one place to another; this is conceptually equivalent to space flowing through the airplane. Instead of fixing the space (i.e., Earth) and moving the plane, we fix the plane and move the space (Earth). In reality, if the airplane were fixed, then when the plane is initially over London, we would find it over New York because the Earth turns around itself. The result is that the plane will be over all cities in its circle around the earth.

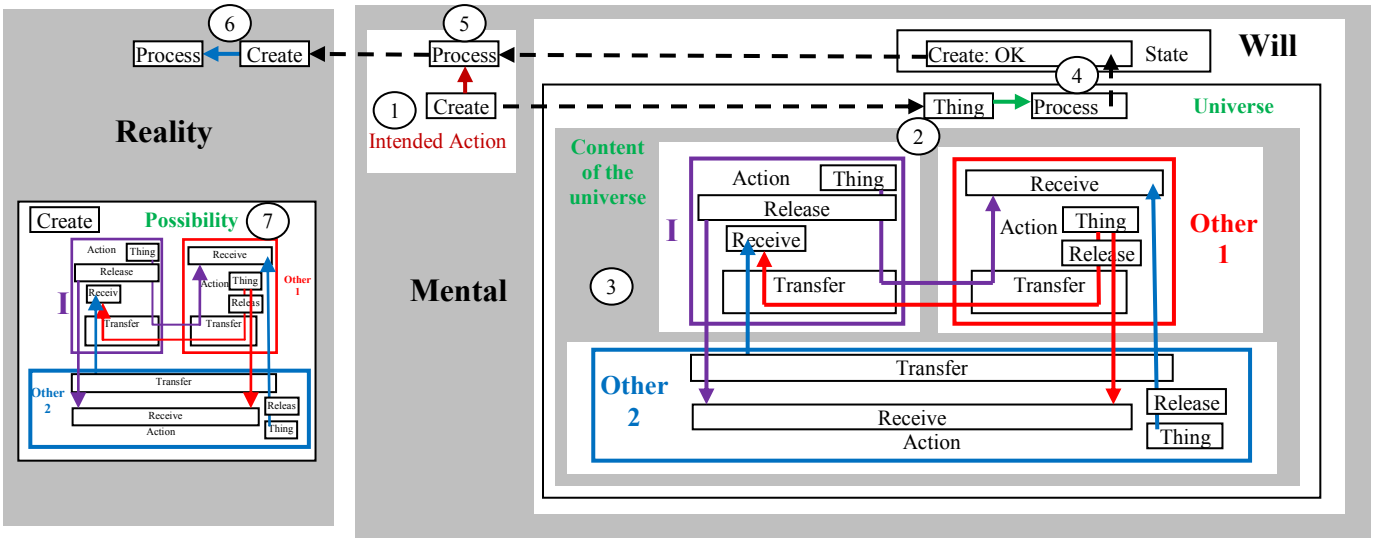


Fig. 12. The TM representation of the categorical imperative.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

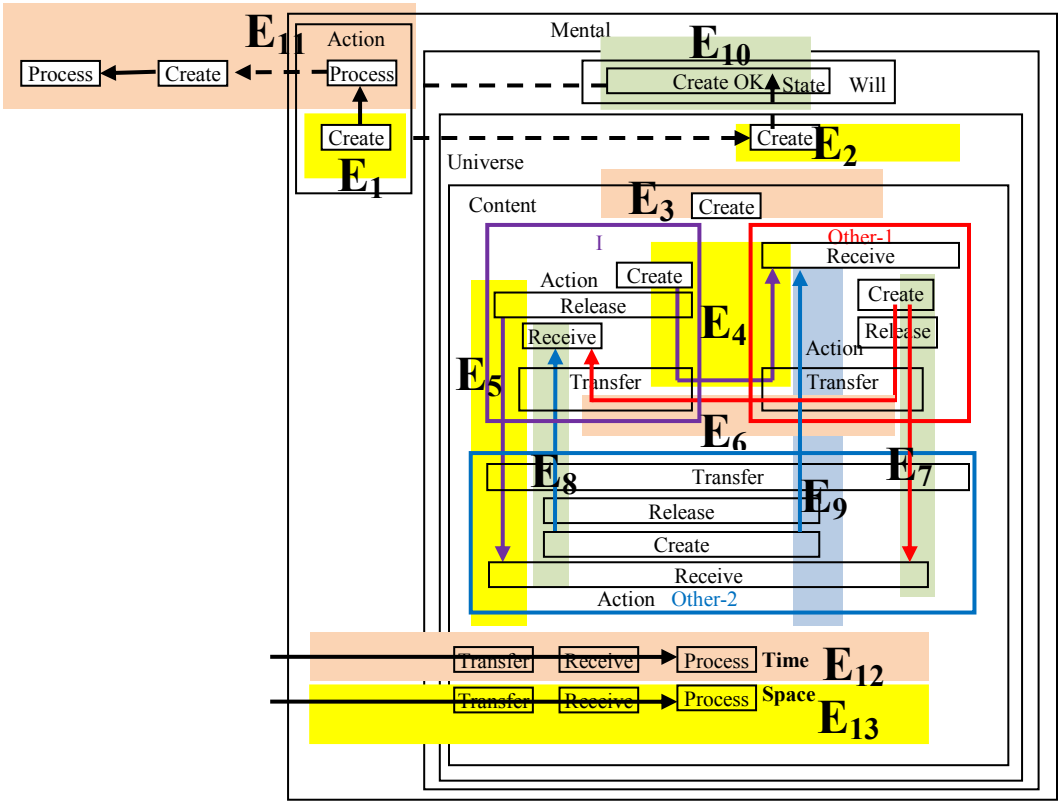


Fig. 13. The TM dynamic representation of the categorical imperative.

Consequently, modeling space (as in the case of modeling time) as flowing in the universe of the figure and never leaving the events means that the events occur everywhere.

Fig. 14 shows the chronology of events, and Fig. 15 illustrates the timing of events wherein some events occur randomly and simultaneously.

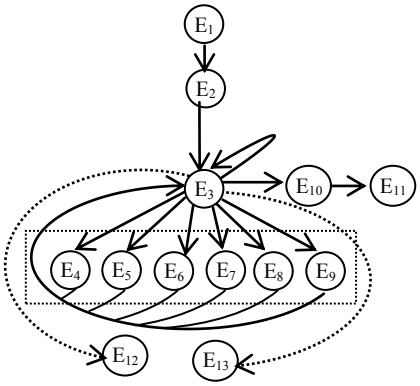


Fig. 14. The chronology of events.

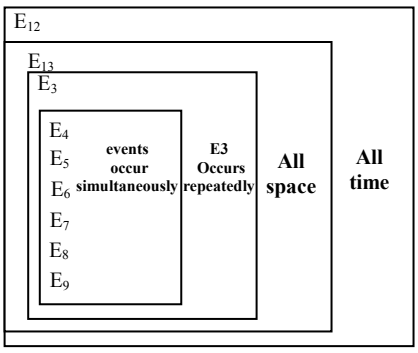


Fig. 15. The timing of events

B. *Kantism and lying*

One of the major challenges to Kant's reasoning is that it is based on the categorical imperative. Since truth-telling must be universal, one must (if asked) tell a recognized killer the place of his prey. According to Kant, it is one's moral duty to be truthful to a murderer. If he/she is untruthful, then this displays a will to end the practice of thinging the truth. The choice in this case is between assisting a murderer and no *wujood* for truth. Lying is fundamentally wrong, and we cannot thing (create) it even when it eventually triggers good. Untruthfulness means willing universal untruthfulness because the net result is that everyone would thing lies. Also, Kant maintained that if a person performs the correct act, telling the truth, then he/she is not blamable for any outcomes.

Fig. 16 models the situation that *one must (if asked) tell a recognized killer the place of his prey*. It includes person 1 (victim—circle 1), person 2 (the murderer—2) and the person who makes an ethical decision (we will refer to him/her as Agent—3). The victim (person 1) hides in a hiding place (4). This is observed by the agent (5). Additionally, the agent observes the character of person 1 as a victim (6 and 7) and person 2 as a murderer (8 and 9).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

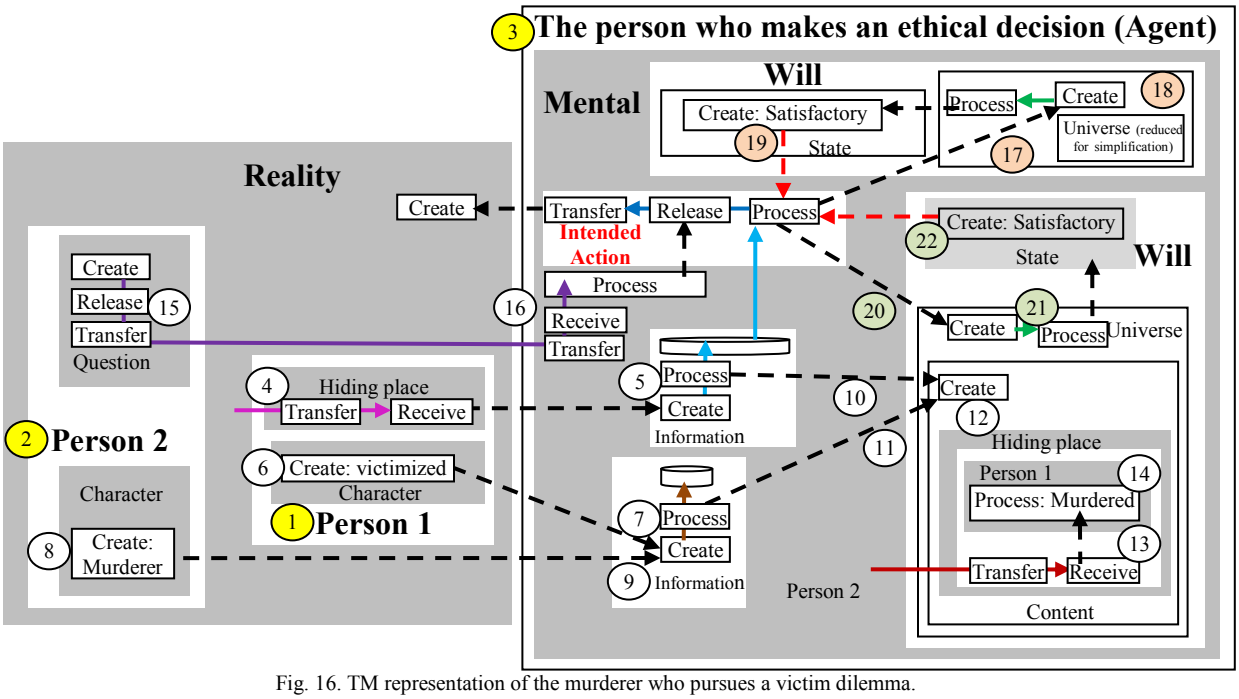


Fig. 16. TM representation of the murderer who pursues a victim dilemma.

From the information about the hiding place (10) and characters of the involved persons (11), the agent creates (12) a mental picture about what will happen if person 2 catches person 1. The picture models the murderer going to the hiding place (13) and murdering the victim (14). Now the murderer asks the agent (15), and the question flows to the agent (16) about the whereabouts of the victim.

Accordingly, the ethical decision to release information about the hideout (15) emerges based on the following factors:

(i) The categorical imperative that triggers the universe (17, 18 and 19), which is "neutral" with respect to the details of the situation. It assumed that the decision is solely based on the categorical imperative. This requires universalizing the act of lying (releasing/transferring misinformation).

(ii) The humanitarian situation as expressed in the mental picture (12). This is triggered by the intended decision (20, 21 and 22) and constructed from the information about the characters of the persons involved (10 and 11). Note that the mental picture appears twice: as imagining what will happen based on information about the characters of the two persons and again as a content of the will. For simplicity's sake, we ignore the appearance of "I" in the universal picture.

Thus, there are two occurrences of universalization: *Everyone is lying* and *every person kills another when given true information about where his/her hideout is*. Thus, the ethical decision to release information about the hideout leads to two contradictory categorical imperatives.

Kant separates individuals from non-individuals when he provides his second formulated principle: the "human integrity" principle. It states: *In every case, treat your own person or that of another, as an end in itself and never merely as a means to an end*. According to Korsgaard [43], "the different formulations [of CI] give different answers to the question of whether if, by lying, someone may prevent a would-be murderer from implementing his/her intentions, that person may do so." From this, it is concluded that different formulations of CI narrow the restrictions imposed by the universalizability requirement.

In our case, we claim that the "human integrity" principle implies that dealing with "information of/about an identified human being" (personal identifiable information) is tantamount to dealing with the human being him/herself.

(iii) There may be other factors, such as the legal issue of assisting a murderer in a crime when telling a recognized killer the place of his prey (not shown in the figure).

The point here is that using TM diagramming facilitates exposing the details of the ethical case. Apparently, the two cases include:

(1) Kant's *pure* categorical imperative as discussed previously, and
(2) Another categorical imperative that involves other details and considerations, as in the case of a murderer pursuing a victim.

Figs. 17 and 18 show different events in this ethical case as follows.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

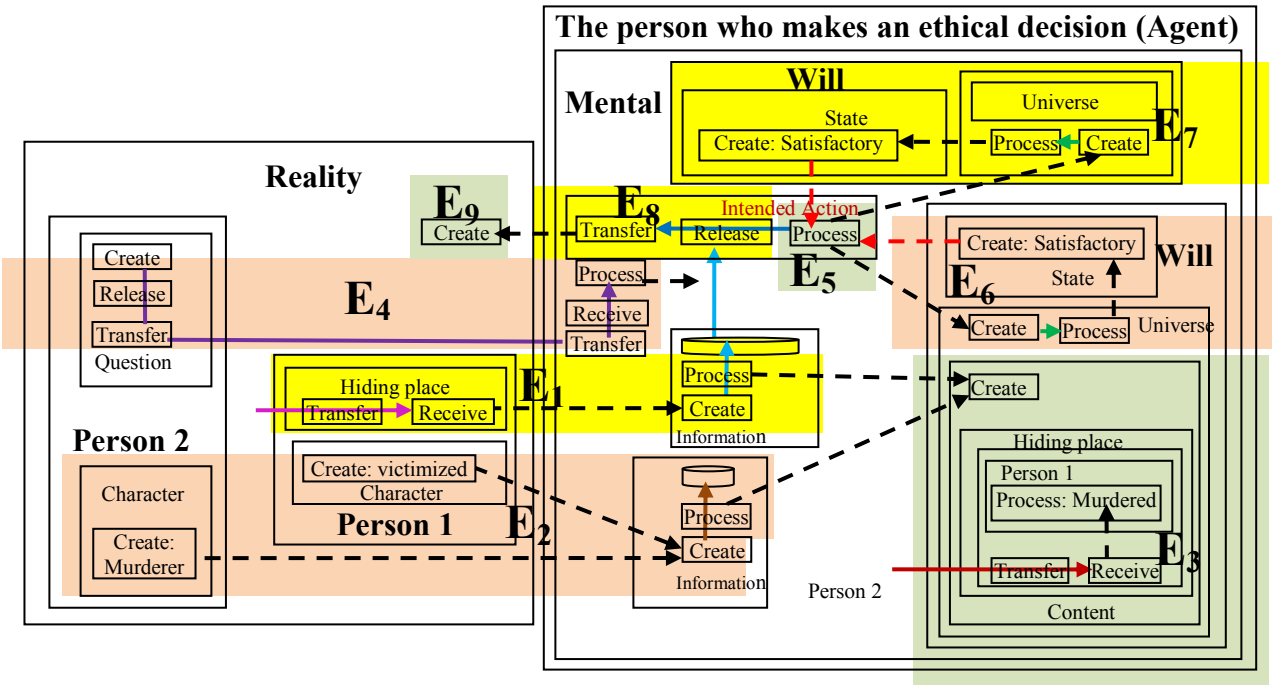


Fig. 17. The TM events of the murderer who pursues a victim.

Fig. 18. The chronology of events.

Event 1 (E_1): The agent observes person 1 (victim) hiding.
Event 2 (E_2): The agent observes the characters of the person to judge that person 1 is a victim and person 2 is a murderer.
Event 3 (E_3): From the information in (2), the agent imagines what will happen if the murderer finds the victim.
Event 4 (E_4): The murderer askes the agent about the whereabouts of the victim.
Event 5 (E_5): The agent processes the intended decision.
Event 6 (E_6): He/she thinks: Do I wish to universalize killing?
Event 5 (E_5): The agent processes the intended decision again.
Event 7 (E_7): He/she thinks: Does he/she will to universalize lying?
Event 8 (E_8): The agent makes a decision about releasing information or misinformation.
Event 9 (E_9): The agent implements his/her decision.

Such a method of modeling ethical decisions is very suitable for software engineers. It facilities a method of diagramming (e.g., flowcharts, UML [11]) that is familiar in their fields. Note that the aim of this paper is to demonstrate the diagramming tool. Thus, if there is some wrong in the ethical thinking (e.g., imprecise understanding of the categorical imperative), then the diagram can be corrected accordingly.

V. Modeling Islamic Ethics

The diagrammatic method described in the previous section has raised the interest of software engineering undergraduate students. However, it is possible to raise their interest further by modeling their own system of ethics. In a multicultural society, different non-secular ethics, such as Islamic, Christian, Judaic, Hindu, and Buddhist ethics can be discussed side-by-side with secular ethics.

This motivates the students because it involves something that is closely related to their personalities. In our school, all students in the computer ethics class are (officially) Muslims. Accordingly, even though there is a lack of knowledge in this area on the part of the instructor, it is possible, with the participation of the students, to develop a reasonably close model of how to make ethical decisions according to Islam. Note that the aim is at using the TM diagrams, so if there is a deviation from the *correct* understanding of Islamic ethics, it easy to modify it.

Additionally, several interpretations of the Islamic ethical system are possible, so selecting the clearest ideas in the literature does not mean adopting these ideas or recommending them for any purpose. This paper discusses an actual experiment in teaching ethics to computer/software engineers with the purpose of exposing them to different approaches and not influencing their ethical thinking.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Accordingly, presenting Islamic ethics was just like presenting Kantism, and selecting Islamic ethics was purely based on the background of the students. For example, in American schools, with multiple students' backgrounds, it is reasonable to model Christian ethics and/or atheist ethics in addition to secular ethics. Note that secularism is not atheism [44].

Theology has a very close relationship with ethics. It includes a history of concern for diverse ethical issues and an important aspect of critical reflection on causes and meanings based on faith and/or a revealed source. Religious ethics are based on divine law. What is right and wrong, what we ought to do or not do, is given by revelation, as moral values and obligations are independent of us. In Islam, moral values and obligations are independent of us; nevertheless, we have complete freedom in selecting to commit ourselves to bringing or not bringing action to the *wujood*.

Actions are judged by intentions, and each action is recompensed according to what a person intends [32].

Fig. 19 shows a model for making an ethical decision in (traditional) Islam according to the understanding of the author. Fig. 20 shows the corresponding event-ized diagram.

In Fig. 19, first a person generates an intended action (circle 1) based on his/her best available information/data and real capability to act (2) and rationality and internal capability to act (3). Then this intended action is processed (4) to check whether it agrees with the Islamic principles as given by the Quran and the Sunna (Sunna is the model pattern of the behavior and mode of life of the Prophet). It is possible that the person might consult an expert or clergy regarding the matter at this stage. Accordingly, the result of this processing is a judgment (fatwa). Either (i) the intended act is permitted in Islam (5) or (ii) it is prohibited by Islam (5).

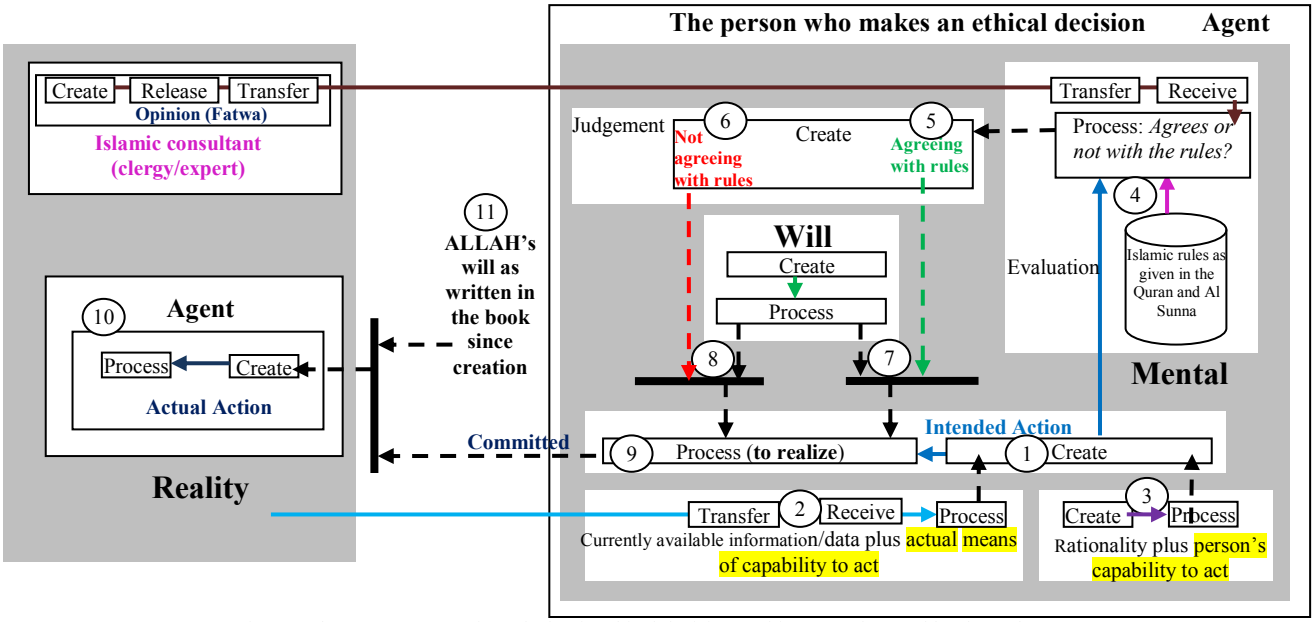


Fig. 19. The TM representation of Islamic ethical decision-making as understood by the author.

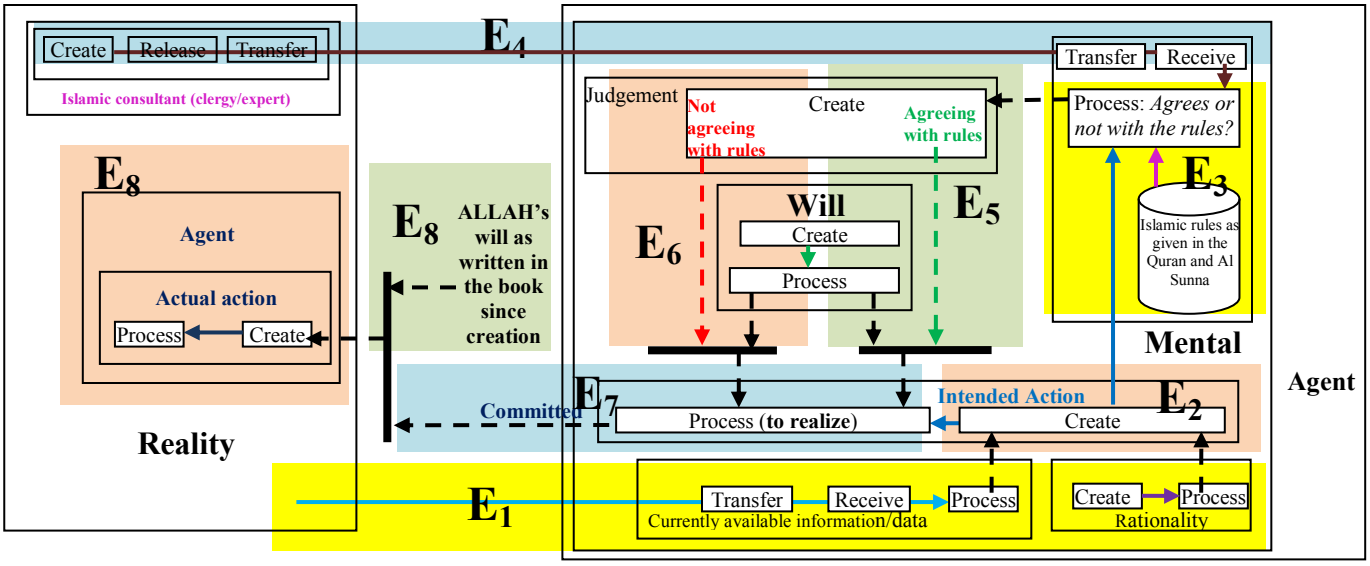


Fig. 20. The TM events of Islamic ethical decision-making as understood by the author.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Then, according to the individual's *free will*, he/she decides to choose which judgement he/she wants to actualize (7 or 8). Consequently, the person "releases" the selected action to actualize it in reality (9). However, in the Islamic faith, all events in reality are written before time (only known by ALLAH) such that no one (even angels) knows what happens until after it has happened. This is called the universal will of ALLAH. Accordingly, an action is actualized in reality (10) only when ALLAH's will (11) coincides with the person's will. This is taken by faith and is an important factor when making an ethical decision as a Muslim.

The set of events of such a scenario (Fig. 20) are as follows.

Event 1 (E_1): The agent receives information/data and reasons for the ethical situation or dilemma.

Event 2 (E_2): The agent creates an intended action.

Event 3 (E_3): The agent applies Islamic rules with regards to his/her intended action.

Event 4 (E_4): The agent may consult an expert in Islam.

Event 5 (E_5): The agent judges that the action is permitted in Islam.

Event 6 (E_6): The agent judges that the action is not permitted in Islam.

Event 7 (E_7): The agent takes and actualizes the judgement from (6) or in (7).

Event 8 (E_8): ALLAH's will is either to actualize such an action or not.

Event 9 (E_9): The action is actualized according to ALLAH's will.

The chronology of these events is shown in Fig. 21.

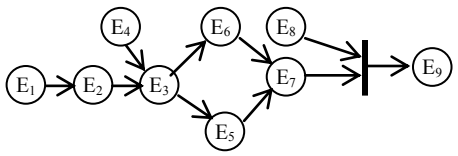


Fig. 21. The chronology of events.

In Kantism, the assumption is that whatever the agent decides by his/her free will is realized in reality; in Islamic ethics, this is not guaranteed because of because of the possible intercession of the will of ALLAH. Note that in Islam, if the an evil action is realized then this does not imply ALLAH's approval (actualizing evil acts) but indicates ALLAH's wisdom in testing humans. If there is no evil in reality then there is no point of free will.

VI. Conclusion

This paper proposes a representational model of ethical systems based on the notions of things and flow. Kantism and Islamic ethics are used as sample ethical systems to demonstrate the applicability of the approach. The modeling technique can be utilized as a pedagogy tool in teaching basic principles of ethics and ethical decision-making. This involves analyzing ethics and design (e.g., robotics). The flowchart-like diagrammatic representation seems to be a familiar style suitable for software engineers.

One weakness of the modeling language is the need to analyze its expressive power of ethical theories and dilemmas that have not been presented in this paper. This is a work for further research.

References

[1] G. D. Crnkovic and R. Feldt, "Professional and ethical issues of software engineering curricula: Experiences from a Swedish academic context," in Proceedings of First Workshop on Human Aspects of Software Engineering (HAOSE09), Orlando, Florida, Oct 25, 2009–Oct 26, 2009. Academic Press.

[2] ACM, ACM Code of Ethics and Professional Conduct, June 22nd, 2018. https://www.acm.org/code-of-ethics

[3] IEEE-CS/ACM Joint Task Force on Software Engineering Ethics and Professional Practices, Software Engineering Code of Ethics, Short version, 1999. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwj46I-Gt8DdAhXChqYKHf30BBEQFjAAegQIABAB&url=https%3A%2F%2Fwww.computer.org%2Fcms%2FComputer.org%2Fprofessional-education%2Fpdf%2Fsoftware-engineering-code-of-ethics.pdf&usg=AOvVaw1_nZEF0dYgHNuiI7iBrJHu

[4] ACM/IEEE Task. Force, "Software engineering code of ethics and professional practice," (version 5.2). https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct

[5] Association for Library and Information Science Education (ALISE). (2007). Position statement on information ethics in LIS education. http://www.alise.org/index.php?option=com_content&view=article&id=51

[6] Status NET, Ethical decision making models and 6 steps of ethical decision making process (No date). https://status.net/articles/ethical-decision-making-process-model-framework/

[7] L. Stickler and G. Sykes, Modeling and explaining content: Definition, research support, and measurement of the ETS® national observational teaching examination (NOTE) assessment series (Research Memorandum No. RM-16-07). Princeton, NJ: Educational Testing Service, 2016.

[8] B. Woodward, Thomas Imboden, "Expansion and validation of the PAPA framework," Information Systems Education Journal (ISEDJ), vol. 9, No. 3, August 2011.

[9] D. G. Johnson, Computer Ethics, 4th edition, Pearson, 2009.

[10] Meherchilakalapudi, Ethical Issues in Software Engineering, blog, March 21, 2009. https://meherchilakalapudi.wordpress.com/2009/03/21/ethical-issues-in-software-engineering/

[11] Object Management Group UML channel, UML & SysML modelling languages: Expertise and blog articles on UML, SysML, and Enterprise Architect modelling tool. http://www.umlchannel.com/en/sysml. [Accessed April 2014]

[12] S. Al-Fedaghi and R. Al-Azmi, "Control of waste water treatment as a flow machine: A case study," 24th IEEE International Conference on Automation and Computing (ICAC'18), 6–7 September 2018, Newcastle University, Newcastle upon Tyne, UK.

[13] S. Al-Fedaghi and M. Bayoumi, "Computer attacks as machines of things that flow," 2018 International Conference on Security and Management (SAM'18), Las Vegas, USA, July 30– August 2, 2018.

[14] S. Al-Fedaghi and N. Al-Huwais, "Toward modeling information in asset management: Case study using Maximo," 4th International Conference on Information Management (ICIM2018), Oxford, UK, May 25–27, 2018.

[15] S. Al-Fedaghi and N. Warsame, "Provenance as a machine," International Conference on Information Society (i-Society), Dublin, Ireland, July 15–18, 2018.

[16] S. Al-Fedaghi and M. Alsharah, "Modeling IT processes: A case study using Microsoft Orchestrator," 4th IEEE International Conference on Advances in Computing and Communication Engineering, Paris, France, June 22–23, 2018.

[17] S. S. Al-Fedaghi, "Thinging for software engineers," International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 7, July 2018.

[18] S. S. Al-Fedaghi and M. Al-Otaibi, "Conceptual modeling of a procurement process: Case study of RFP for public key infrastructure," International Journal of Advanced Computer Science and Applications (IJACSA), Vol 9, No 1, January 2018.

[19] S. S. Al-Fedaghi, "Privacy things: Systematic approach to privacy and personal identifiable information," International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 2, February 2018.

[20] S. Al-Fedaghi and H. Aljenfawi, "A small company as a thinging machine," 10th International Conference on Information Management and Engineering (ICIME 2018), University of Salford, Manchester, United Kingdom, September 22–24, 2018.

[21] S. Al-Fedaghi and M. Alsharah, "Security processes as machines: A case study," Eighth international conference on Innovative Computing Technology (INTECH 2018), London, United Kingdom, August 15–17, 2018.

[22] J. Carreira, "Philosophy is not a luxury," [blog], March 2, 2011, https://philosophyisnotaluxury.com/2011/03/02/to-thing-a-new-verb/

[23] M. Heidegger, "The thing," in Poetry, Language, Thought, A. Hofstadter, Trans. New York: Harper & Row, 1975, pp. 161–184.

[24] Joël de Rosnay, "The macroscope," Principia Cybernetica Project, Translated by Robert Edwards, Harper & Row, 1979.

[25] K. K. Humphreys, What an Engineer Should Know About Ethics, New York/Basel: Marcel Dekker, Inc., 1999.

[26] J. Ladd, "The quest for a code of professional ethics: An intellectual and moral confusion," in Deborah G.Johnson, Ethical Issues in Engineering, Prentice-Hall, Engelwood Cliffs, 1991, pp. 130-136.

[27] M. W. Martin and R. Schinzinger Ethics in Engineering, Third Edition. New York: McGraw-Hill, 1996.

[28] S. D. Ross, Ideals and Responsibilities: Ethical Judgment and Social Identity. Belmont, CA/Albany NY: Wadsworth Publishing Company, 1998.

[29] P.A. Facione, D. Schere, and T. Attig, Ethics and Society, Englewood Cliffs, NJ: Prentice-Hall, 1991.

[30] A. Sesonske, Value and Obligation: The Foundation of Empiricist Ethical Theory. New York: Oxford University Press, 1964.

[31] B. Rosen, Ethical Theory. Mountain View, California: Mayfield Publishing Company, 1990.

[32] G. F. Hourani, Reason and Tradition in Islamic Ethics. Cambridge/London/ New York: Cambridge University Press, 1985.

[33] P. Edwards, Encyclopedia of Philosophy, Vol. 5, New York: Macmillan, 1967.

[34] C. G. Christians, "Ethical theory in a global setting," in Communication Ethics and Global Change, T. W. Cooper (general editor), New York: Longman Inc., 1989.

[35] M. E. Clark, Ariadne's Thread: The Search for New Modes of Thinking. London: Macmillan Press Ltd., 1989

[36] M. Pascal, My Philosopher: Immanuel Kant, 2015, accessed August 2018. https://mediaethicsafternoon.wordpress.com/2015/02/13/my-philosopher-immanuel-kant/

[37] Jürgen Habermas, Stanford Encyclopedia of Philosophy, 2014. https://plato.stanford.edu/entries/habermas/

[38] John Rawls, Theory of Justice, Cambridge, Mass.: The Belknap Press of Harvard University Press, 1971.

[39] L. McKnight, Immanuel Kant and "the Categorical Imperative" for Dummies, Owlcation, 2016.https://owlcation.com/humanities/Immanuel-Kant-and-The-Categorical-Imperative

[40] Moral Robots, Kant's Ethics, Blog. 2017. https://moral-robots.com/philosophy/briefing-kant/

[41] S. McLachlan, K. Dube, T. Gallagher and J. A. Simmonds, "Supporting preliminary decisions on patient requests for access to health records: An integrated ethical and legal framework," 2015 International Conference on Healthcare Informatics (ICHI), Dallas, TX, USA, Oct. 21–Oct. 23, 2015.

[42] G. Marckmann, H. Schmidt, N. Sofaer and D. Strech, "Putting public health ethics into practice: A systematic framework," Front Public Health. Frontiers in Public Health Vol. 3, No. 23, 2015. doi:10.3389/fpubh.2015.00023.

[43] C. Korsgaard, "The right to lie: Kant on dealing with evil." Philosophy and Public Affairs vol. 15, no. 4, 1986.

[44] J. Berlinerblau, Secularism Is Not Atheism, The Blog, Dec 06, 2017.https://www.huffingtonpost.com/jacques-berlinerblau/secularism-is-not-atheism_b_1699588.html

Critical Analysis of High Performance Computing (HPC)

Misbah Nazir, Dileep Kumar, Liaquat Ali Thebo,
Syed Naveed Ahmed Jaffari,
Computer System Engineering Department,
Mehran University of Engineering & Technology,
Sindh, Pakistan.

kmisbah90@yahoo.com,
dileeplohana.engr@gmail.com,
liaquat.thebo@faculty.muet.edu.pk,
Naveed.jaffari@faculty.muet.edu.pk.

***Abstract*—Parallel computing has become most important issue right this time but because of the high cost of supercomputer it is not accessible for everyone. Cluster is the only technique that provides parallel computing, scalability and high availability at low cost. Collection of personal computers (PCs) builds a cluster that provides us parallel execution. High Performance Computing (HPC) is the field of computer science that emphases on making of cluster computers, supercomputers and parallel algorithms. At this present time, clusters technique has practical in numerous areas, for instance scientific calculations, weather forecasting, bioinformatics, signal processing, petroleum exploration and so on. This paper compares the two different types of clusters to check the overall performance in execution time. One cluster is made up of Dell core 2 duo systems and second cluster is made up of HP core 2 duo systems each with two nodes having almost same configurations. To analyze the performance of these two clusters we have executed two different parallel programs on the clusters for pi calculation and quick sort with different problem sizes. We observed that with small size problems Dell cluster performed better against HP cluster while with large size problems HP cluster won the game.**

Keywords-component; Parallel Computing, Beowulf Cluster, High Performance Computing.

I. INTRODUCTION

Parallel computing has become most important issue right this time but because of the high cost of supercomputer it is not accessible for everyone. Cluster is the only technique that provides parallel computing, scalability and high availability at low cost. Collection of cluster provides high performance computing, independent computers work simultaneously to solve a problem. Generally, machines are coupled at one place, which is connected by network. The main goal of HPC is to crunch numbers, rather than to sort data. It requires special program optimizations to acquire the maximum from a system in the form of input/output, computation, and data movement. The machines all have to confidence each other because they are sending information to and fro. Collections of personal computers (PCs) build a cluster that provides us parallel computing. Every node in a cluster share processor(s) and other multiple resources with each other to analyze and compute the complex computational problems and work as a single machine. As we know the cost of a supercomputer is millions of dollars but now in these days cluster computing with Linux has overcome this problem. Powerful, stable and efficient cluster can be created in low budget. These clusters can add any number of nodes and provide HPC environment to replace supercomputers.

For parallel processing HPC Cluster use software called parallelized software, so in this manner, problem divided into chunks and distributed over a network nodes which are interconnected so input can be process and communicate for final output. The biggest advantage of cluster computing is that Cluster nodes take less space, less cooling, less power and low maintenance cost. The favorable condition and advantage of such cluster setup is quick solution of complex jobs by dividing into smaller parts and it's execution through parallel processor. In such a manner HPC can be achieved in different manner using cluster approach but still there are many challenges remaining to overcome.

Clusters are designed because they provide high performance computing and availability over single computer. A cluster is a group of connected devices such as computer and switches and working together as a single system. Each and every node of a cluster is associated with each other either by wire (Ethernet) or wireless that pass data between the nodes. A Beowulf cluster provides distributed computing. It is made of standard desktop computers that are linked via network such as Ethernet. Linux operating system can be used to control the cluster. In this way, we can achieve high performance computing (HPC) at low-cost price.

This project compares the two different types of clusters to check the overall performance such in execution time. One cluster is made up of using two Dell core 2 duo systems and second cluster is made up of using two HP core 2 duo systems. Both have almost similar configuration. The primary goal of this research is to compare the execution time difference between these two clusters and analyzes the performance of the clusters. A consequence of providing such environment is that we can find out which cluster formation performs more efficiently. The focus of this research is the cluster computing systems, which are a specific type of computer clusters developed primarily for computational purpose.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The remaining part of the paper is ordered in sections as: Section II enlightens the background of cluster computing. Section III portrays cluster computing. Section IV describes the methodology. Section V illustrates the design of cluster. Section VI provides testing performance and results and finally section VII concludes the paper and section VIII suggests the future work, followed by references.

II. BACKGROUND

In this paper author used 4 Raspberry Pi computers to create low cost cluster and compared the cluster with single Raspberry Pi computer and he found that the cluster computing are more useful in large and complex problems rather than small and low computational problems[1,2,3]. This paper presented a physical cluster for teaching cloud computing and big data. He has presented three different types of configuration of physical cluster [4]. In this research, they built a cluster of using 4 raspberry pi 2 computer which is basically a multicore computer and compared the model B of Raspberry pi 2 with quad core operating at 0.9 GHz with older model of Raspberry pi with single core operating at 0.9 GHz. The results are shown in the statistical form [5, 6, 7, 8]. This paper discussed high performance computing established on Linux cluster. Author used cpi algorithm to check the performance of cluster. He analyzed the performance by using single and multiple nodes by using different intervals [9,10]. Author designed a low cost Beowulf cluster for academic purpose. Undergraduate and postgraduate students of computer science can use that cheap cluster in parallel and distributed computing courses. To analyze the performance of the cluster, researcher executed two parallel programs on a cluster with different number of nodes [11]. Research compares and examines the performance between cluster and multicore architecture processor. He used 14 raspberry pi devices to build a cluster and run two programs to check the performance on the basis of CPU time in FLOPS (Floating point operations per second) and run same program on core i5 and core i7 processor for the same purpose [12,13]. The aim of the research is to design a Beowulf cluster to solve the complex computational problems. He used Linux operating system and implemented MPI (message passing interface) technique developed in C language [14,15]. Here author designed an algorithm to compare the performance between two multicore processors (kepler GPU processor and Xeon Phi processor) which is specially designed for high performance computing. He used OpenCL environment to implement algorithm [16]. Aaron designed a cheap cluster for HPC for educational area. He used 25 Raspberry pi model B computers to build a cluster. One is master node and rest of the slave nodes. Head node sends the computational task to all worker nodes. He used Raspbian operating system which is type of Linux operating system and python language for testing. He written three test in python code that demonstrate the linear algebraic operations between matrices, scalar and vector [17, 18, 19, 20]. The project described the prototype model of Beowulf cluster. Beowulf cluster is built on Linux operating system. TCP/IP protocol used to implement MPI library. This protocol layer passes the messages along the operating system onto a physical network. The cluster is made of heterogeneous personal computers (PCs) [21, 22, 23]. Author designed and implement cost effective HPC cluster for teacher oriented computer science curriculum. Cluster based on Play station 3 (PS3) which is a game console and each PS3 consider as a single processing node. Cluster consist of six PS3 nodes and install fedora core 12 OS on each node [24, 25, 26]. This project designed the cluster system consist of six personal computers (PCs) and tested the performance of the cluster by using High performance LINPACK (HPL) software [27, 28]. This project proposed a cost effective self-training algorithm for load balancing of every node in cluster computing and compared their proposed algorithm with another traditional load balancing algorithm. Author used 16 nodes to build a cluster. All computing task ran Discrete Fourier transform [29,30,31]. Multi-core architecture provides more scalability than single core architecture in cluster computing. Author performed some experiments to check the impact of multicore architecture on cluster computing. High performance LINPACK (HPL) software used as a benchmark in Intel Bensley system. He also compared the multicore cluster with the single core cluster and conclude that single core cluster have same scalability as the multicore cluster [32]. In this paper, author built PCI Express (Peripheral Component Interconnect *Express)* based cluster using COTS (commodity off-the-shelf) technologies and evaluated their outcomes with SGI Altix 3700. For the flexible implementation of parallel sparse linear algebra operations obtained results shown that the most critical factor between the nodes in a cluster is bandwidth [33, 34, and 35]. This research proposed a parallel hierarchical clustering algorithm named PARC. Firstly author analyzes theoretically data parallelism in PC cluster and experimental results obtained from PARC algorithm proved the correctness of their theoretical analyses [36, 37, 38,39]. In this paper author designed the cluster for the researchers which consist of five nodes (one is front-end node and rest are computational nodes). Only front end node has a connection to the world through the internet [40, 41, 42].

III. CLUSTER COMPUTING

A supercomputer is one of the largest and fastest computers at this time. So, the definition of supercomputing is constantly changing. Supercomputing is also called High Performance Computing (HPC). At this present time, clusters technique has practical in numerous areas, for instance scientific calculations, weather forecasting, bioinformatics, signal processing, petroleum exploration and so on.

Cluster is the only technique that provides parallel computing, scalability and high availability at low cost. Collection of cluster provides high performance computing, independent computers work simultaneously to solve a problem. Generally, machines are coupled at one place that is connected by network. Collections of personal computers (PCs) and networked devices build a cluster that provides us parallel computing. Every node in a cluster share processor(s) and other multiple resources with each other to analyze and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

compute the complex computational problems and work as a single machine. To transfer the data between the nodes, the nodes of a cluster are usually connected via wires (Ethernet cable) but not always.

A. Reason to use Cluster

Despite the fact that clusters can be implemented on different operating systems like Solaris, Windows, Macintosh, etc. To use Linux operating system in cluster development has some advantages as follows:

- Linux supports extensive variety of hardware.
- Linux is excellently stable.
- Linux is open source and distributed without any cost.
- Linux is comparatively virus free.
- Linux has broad mixture of tools and programs freely available.
- Excellent for creating clustering systems.

The main reason to develop a cluster instead of using a single computer is that it provides better performance and availability. To achieve better performance at low cost, Beowulf cluster is the best choice. Scientific application require greater performance, this is another important reason for the development of Beowulf clusters through which we can achieve HPC at low cost. Some applications require more computational power for many reasons, but the most common three reasons are described below:

1) The need for Real-time constraints

The main requirement is the computation of a process must be completed within dedicated time for instance weather forecasting. A further important factor is execution; the data must be processed as fast as possible.

2) Increase of Throughput

Sometimes single processor would take couple of days and years to compute a single computational problem because that problem requires so much computing power. The cluster is a cost effective technique that increase the computing power. Cluster made it easy for us to solve complex and large number of computational problems easily mostly in scientific and engineering areas.

3) Memory capacity

Some applications require very large amounts of data. Cluster technology provides an effective way of program memory in terabytes for those applications. As fault tolerance is very much important for assuring the availability of computational power to system and that is provided by cluster. Parallel programming helps to achieve high computational power in clusters. Parallel programming is a technique in which multiple processors coordinate with each other to solve a single problem.

4) Cluster Framework.

Cluster computing is a technique in which multiple computers are connected together, so in that way they work as a single machine. A cluster can consist of number of nodes (at least two) which have their own data space. When an application runs on any node, the data of the application have been stored on the data space that node shared. Application program file and operating system of each node are stored in their local repository. Network provides communication between every node of a cluster. Whenever a fault occurs on any node, the running program with this node will be in use over another node. Fig.1 describes the basic framework of cluster.

There are so many advantages of implementing the cluster architecture for instance high availability, scalability, highly usability, and so on, which solve the large scale scientific and engineering application problems.

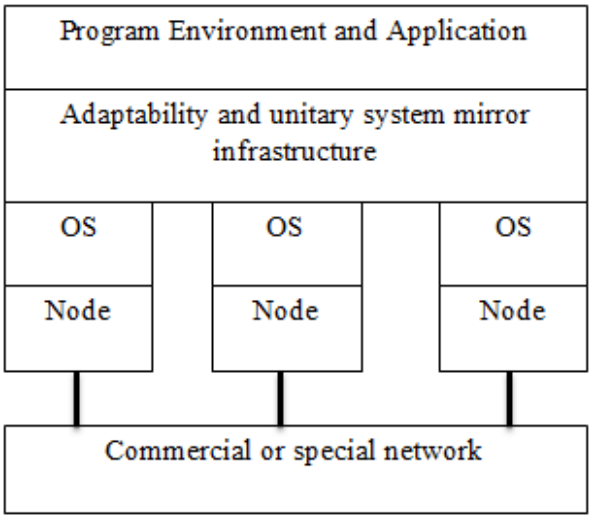


Figure 1.Framework of Cluster

IV. Methodology

Collections of personal computers (PCs) build a cluster that provides us parallel computing. Every node in a cluster share processor and other multiple resources with each other to analyze and compute the complex computational problems. There are many other factors and components are used to create a cluster that works together i: e operating system, hardware consideration, and networking. Here is the view of those components that are important to create a cluster.

A. Cluster Components

To manage the cluster as single machine Ubuntu operating system and the interface software are used. These softwares are installed on every node of the cluster (such as master node and slave node). The main controller of the cluster is master node where full operating system would be installed and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

configures in order to control and supervise all the slave nodes. The slave nodes are considered as computed node.

B. *Operating system*

The operating system is used in this project is Ubuntu operating system with the version 16.04. The main reason for using Ubuntu operating system is its pliability and adaptability because Ubuntu operating system is open source software. Ubuntu is the software on which we performed operations on the cluster. The kernel of Ubuntu operating system handling and supervisory the resources and different parts of the cluster and allocating the jobs to the nodes and also obtaining the outcomes from the nodes and output is a solitary.

To design the environment of parallel computing, we chose two personal computers in each cluster, decided one of them as master node. The second is considered as slave node computer. All of these PCs are connected through switch and created star shaped network. Ubuntu operating system and all necessary tools package we installed on each computer.

C. *SSH (Secure shell)*

Secure Shell is an application to access other computers through network. SSH executes commands in a system and also provide a facility to transfer files from one computer to a different computer. It offers tough authentication and secure interactions over insecure channels. The Secure Shell is not installed automatically with Ubuntu, it should be downloaded.

After the SSH has been installed, the user has the right to use several of the nodes within the cluster. When a user needs to duplicate configuration files over multiple nodes of the cluster, SSH provides this ability and it is very useful.

D. *NFS (Network File system)*

NFS stands for Network File System. It is a distributed file system which is developed by Sun Microsystems, Inc. NFS is a client/server architecture. NFS allows users to access files through a network and entertain that files as they are existing in their local repository. Let's take an example, if you are using a computer that are connected with another computer through NFS, so you can access the files in the second computer even though these file are resided in the repository of first computer. NFS not only provide facility to share files over a network even it's also provide a facility to share resources between multiple nodes.

E. *MPICH2 Installation and configuration*

MPICH2 is a portable implementation of MPI (Message Passing Interface). MPI permits processes to correspond with each other by transmitting and getting messages between them. It is usually utilized for creating parallel programs and executing on cluster computers and supercomputers. Processes can communicate with each other by sending and receiving messages because each process has separate address space. Every process would be processed on a different processor. MPI is moreover supporting shared-memory architectures. This implies that several processes can read or write to the similar memory location.

MPICH2 is Free Software and is available in most of the flavors of Microsoft Windows and UNIX which is providing the separation between communication and a process management. There is a set of daemons called mpd (multi-purpose daemon) present in default runtime environment which is responsible for establishing the communication requires at starting of the application process between the machines which results showing of clear image of wrong happening when so ever communication could not be able to establish hence its providing a scalable and fast startup mechanism once parallel jobs are started.

F. *Hardware Consideration*

Installing the software is essential to access the computers in cluster. Therefore, the installation of the software is held at the early stages.

For building a cluster, it is essential to have at least two computers but it is not essential that these computers have same configuration. The only condition is that they both share same architecture. For example, the clusters should contain either all Intel computers or all Apple computers but not a mix of them. The major hardware necessities for developing a cluster are, at-least two computers and a networking switch to connect with them.

V. DESIGN THE CLUSTER

Clustering technique can be implemented by using different operating system but here we are using Ubuntu operating system which is the distribution of Linux. To design Linux computer cluster we require collection of personal computers that provides parallel computing. Cluster is the only technique that provides parallel computing, scalability and high availability at low cost. Collection of computers provides high performance computing, independent computer work simultaneously to solve a problem. Generally, machines are coupled at one place, which is connected by network. Every node in a cluster share processor(s) and other multiple resources with each other to analyze and compute the complex computational problems. Ubuntu operating system is used to run the cluster as single machine. This OS is installed on master node and slave nodes of the cluster. The master node is a controlling node through which user can login. The slave nodes consider as a computational nodes.

The cluster of two nodes is networked together using switches and cables and formed star shape LAN structure. The following flowchart describes the important steps taken to build a cluster which is shown in figure 2.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

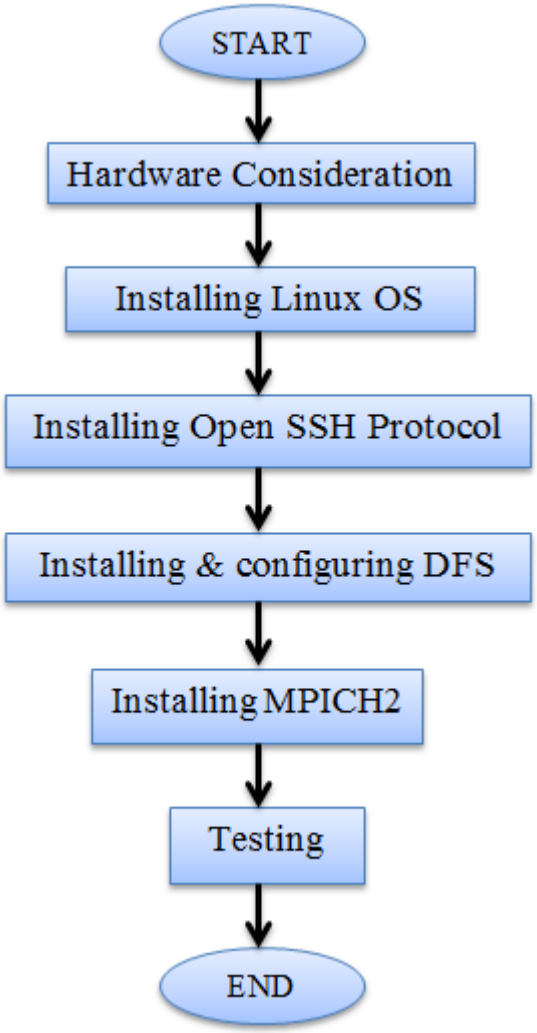


Figure 2.Steps of Building Computer Cluster

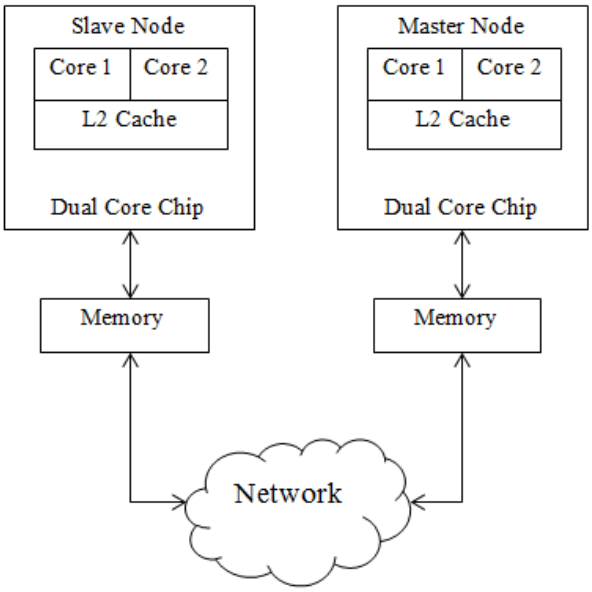


Figure 3(a). Block Diagram of Dell Cluster

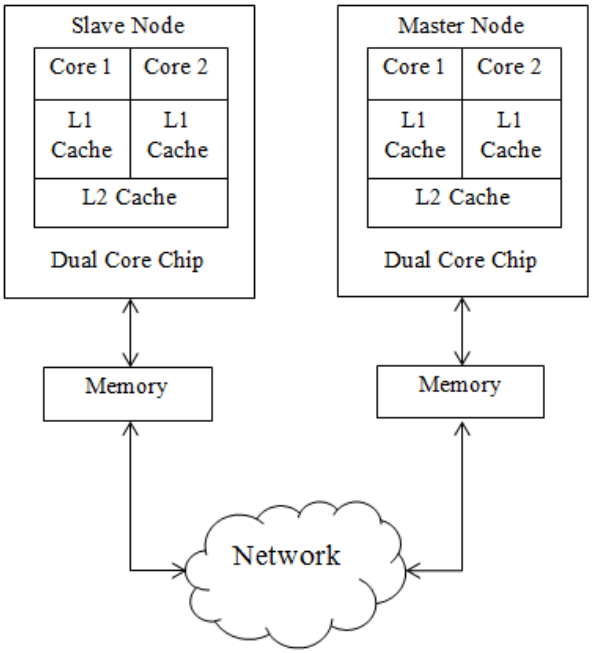


Figure 3(b). Block Diagram of HP Cluster

A. System Design

Here we designed a cluster by connecting two nodes that are connected by network with each other and act as a single machine. MPI can be installed on the computer after the completion of networking. MPICH2 is software through which we can send and receive messages between nodes. It allows the cluster to work as a single united computing unit. It is shown in figure 3(a) and 3(b) that describes the multicore processor architecture of Dell and HP cluster. It can be observed from the diagram that dell and hp processors are dual core architecture. It consists of 2 GB RAM and processor speed is 2.33GHz. The one and only difference between these two architectures is that HP processor contains L1 and L2 cache whereas Dell processor contains only L2 cache.

VI. Testing Performance and Results

To obtain the results as well as to test the performance of HPC cluster at first we have to run or compile the two programs on the master node of the cluster.

A. Calculating Value of Pi

A program to calculate the exact value of mathematical constant PI (3.14) was assessed for passed time and error in the calculated value. The error was perceived to show very small change which negligible and therefore we concentrated

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

mainly on the execution time of the program. Immense use of machine file was made for submitting the processes.

The processes are permitted to move back and forth the master node depending upon the free resources. For instance, master node was allowed two processes at a time. When the five processes are to be scheduled, two will be scheduled on master node and other two will be scheduled on the slave node. At the same time the fifth process will be queued and meanwhile one of the processes inside the cluster is terminated. Therefore, processes were distributed dynamically depending on the free resources on a node, it could be a master node or slave node.

The following tables showing the execution time of HP cluster and Dell cluster after compiling the pi program with single and multiple processes by using different number of intervals as input.

TABLE I
CALCULATED EXECUTION TIME OF HP CLUSTER

No: of Processes	Average execution time with different number of intervals		
	10 Million	100 Million	200 Million
P1	3.1030707	30.9990884	61.9938642
P2	1.5563503	15.5059589	31.0026928
P3	1.0583316	10.3770208	20.7131414
P4	0.7972842	7.7988261	15.6195543
P5	0.9758404	9.4295536	18.7894816
P6	0.8523526	7.9726716	15.8272583
P7	0.9114953	8.9134326	17.381968
P8	0.806848	7.8257761	15.5530288

TABLE II
CALCULATED EXECUTION TIME OF DELL CLUSTER

No: of Processes	Average execution time of different number of intervals		
	10 Million	100 Million	200 Million
P1	3.0970974	30.9603916	61.9163306
P2	1.5507987	15.4847648	30.962437
P3	1.0459178	10.3648435	20.6869724
P4	0.7848046	7.7748677	15.5403275
P5	0.9897083	9.6023393	19.1414227
P6	0.846881	8.1635569	16.0203394
P7	0.9004311	8.9092722	17.7684196
P8	0.7958912	7.7822047	15.5259615

The information given in the table I and table II represents average execution time of different number of intervals from process 1 (P1) to process 8 (P8). It can also be represented by graphs to analyze and understand the performance more easily through graphical representation.

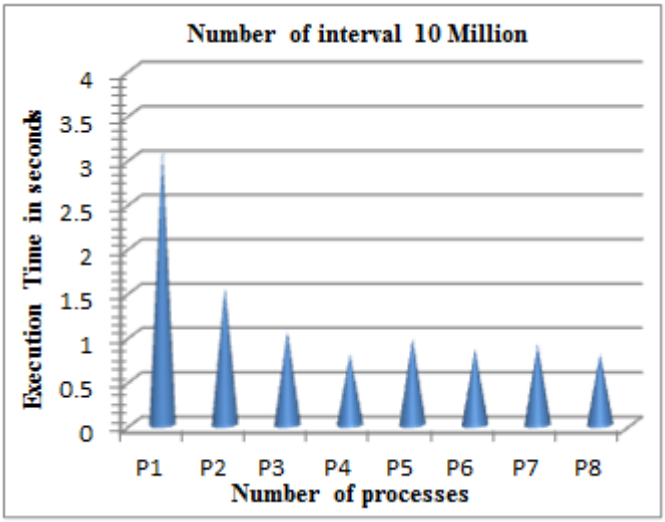


Figure 4.Graphical representation of HP Cluster (10 million Intervals)

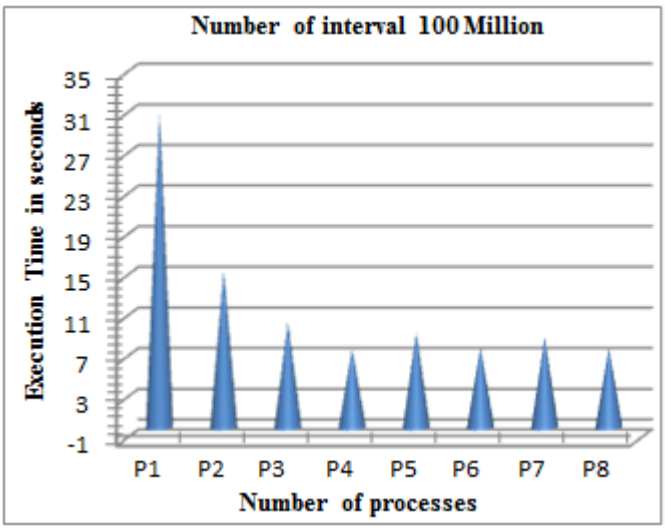


Figure 5.Graphical representation of HP Cluster (100 million Intervals)

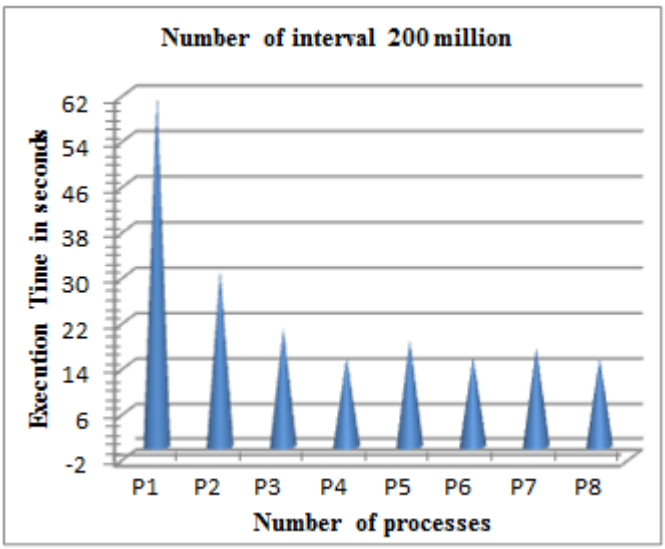


Figure 6.Graphical representation of HP Cluster (200 million Intervals)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

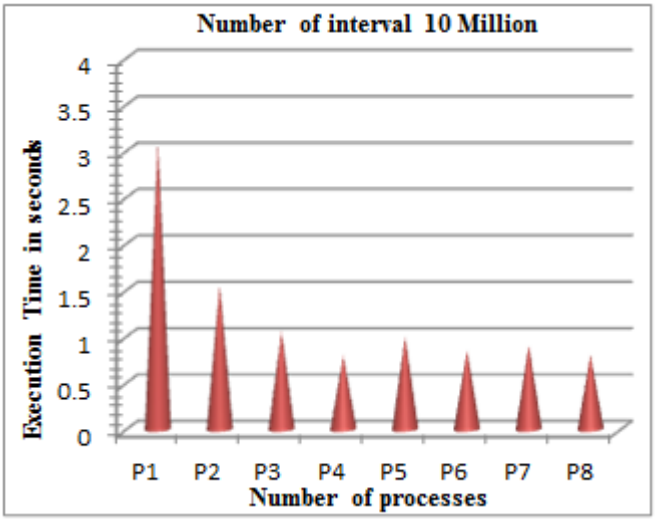


Figure 7.Graphical representation of DELL Cluster (10 million Intervals)

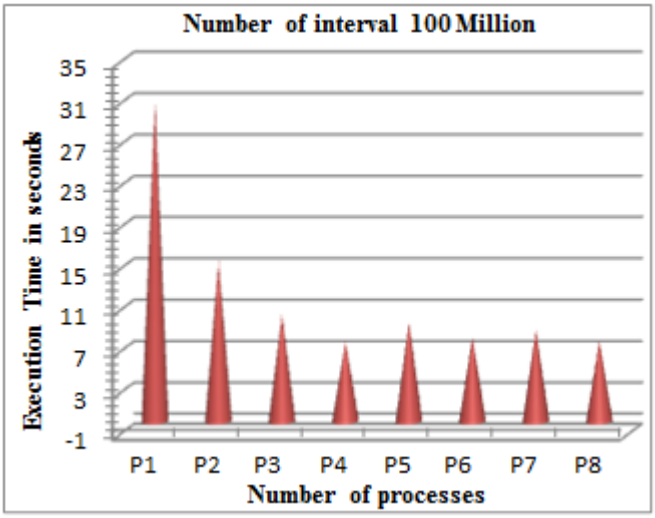


Figure 8.Graphical representation of DELL Cluster (100 million Intervals)

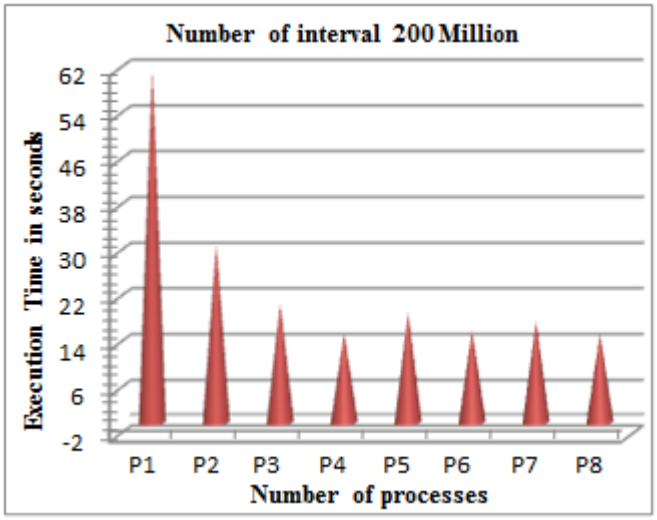


Figure 9.Graphical representation of DELL Cluster (200 million Intervals)

From above all graphical representations of HP cluster and Dell cluster figures (4 to 9) showing the execution time with different number of intervals as Problem size is proving that serial execution takes longer time than parallel execution and also as the number of processes increases, the execution time also decreases. As single process (P1) is taking longest execution time whereas P4 is taking shortest execution time amongst all P1 to P8. It is also analyzed that execution time of the same problem increases after P4 because 4 processes can run simultaneously at the same time on the cluster and other processes must wait until the completion of anyone process.

1) Time Difference Between Dell and HP Cluster

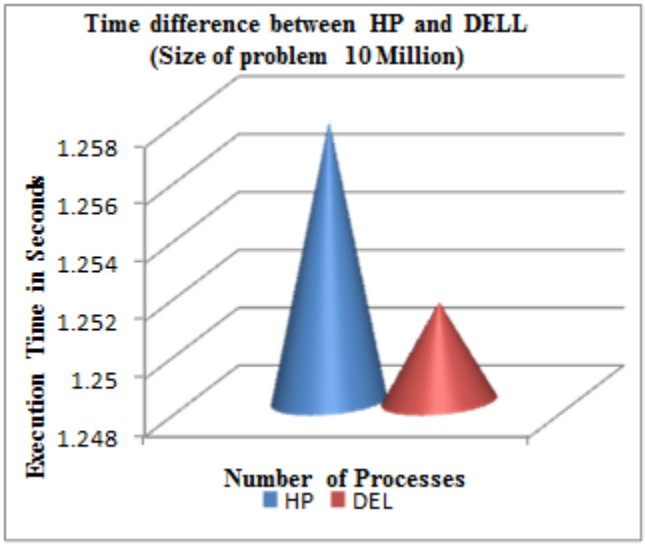


Figure 10.Execution Time Difference between Dell and HP (10 million Intervals)

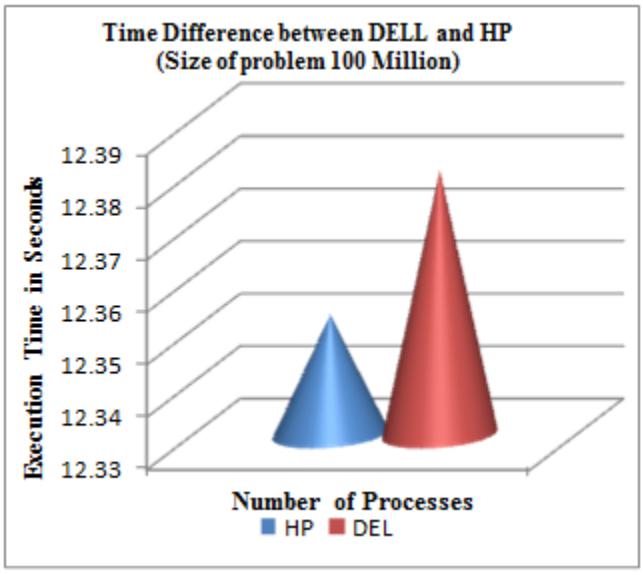


Figure 11.Execution Time Difference between Dell and HP (100 million Intervals)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

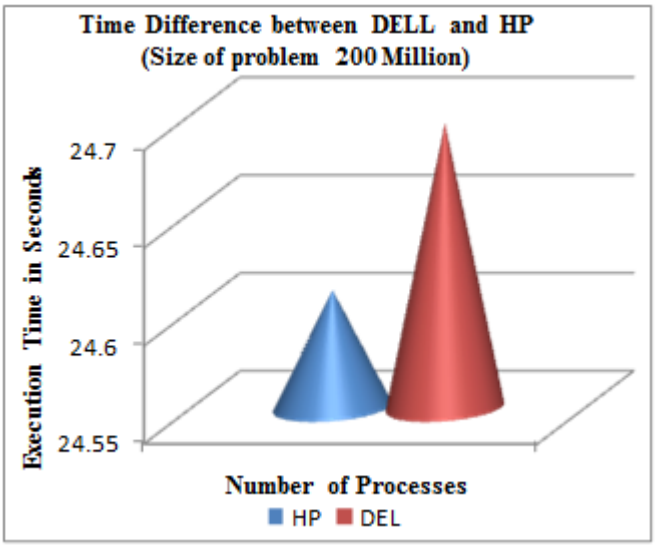


Figure 12.Execution Time Difference between Dell and HP (200 million Intervals)

It can be observed from figure 10, the Dell cluster executes the pi problem 6 milliseconds before the HP cluster while in the figure 11 and 12, it shows clearly that the HP cluster provide better performance. It executes the pi problem 27 milliseconds and 85 milliseconds before the dell cluster. So here we conclude that for the small size of problem, Dell cluster provides better performance than HP cluster but when increasing problem size, HP cluster provides better performance than Dell cluster.

B. Sorting Random Numbers

The second algorithm which we run on the cluster was sorting the random numbers which is saved in input files. There are different sizes of files which are executed on cluster. The following table showing the execution time of HP cluster and Dell cluster after compiling the sorting program using single and multiple processes with different input sizes.

TABLE III
CALCULATED EXECUTION TIME OF HP CLUSTER

No: of Processes	Average execution time with different input size		
	1 Thousand	10 Thousand	100 Thousands
P1	0.0017	0.1522	1532.4611
P2	0.0035	0.0146	1.436
P3	0.0042	0.0129	1.0323
P4	0.0054	0.0151	1.1572
P5	0.0068	0.0159	1.0334
P6	0.0089	0.0171	1.1247
P7	0.0091	0.016	1.0569
P8	0.0108	0.0176	1.0642

TABLE IV
CALCULATED EXECUTION TIME OF DELL CLUSTER

No: of Processes	Average execution time with different input size		
	1 Thousand	10 Thousand	100 Thousands
P1	0.0018	0.1514	14.8986
P2	0.0032	0.0137	0.1279
P3	0.004	0.0117	0.0973
P4	0.0053	0.0143	0.1108
P5	0.007	0.0156	0.1032
P6	0.0079	0.0171	0.1008
P7	0.009	0.0177	0.1009
P8	0.0099	0.0175	0.1151

The information given in the table III and table IV represents average execution time of different input sizes from process 1 (P1) to process 8 (P8). It can also be represented by graphs to analyze and understand the performance more easily through graphical representation.

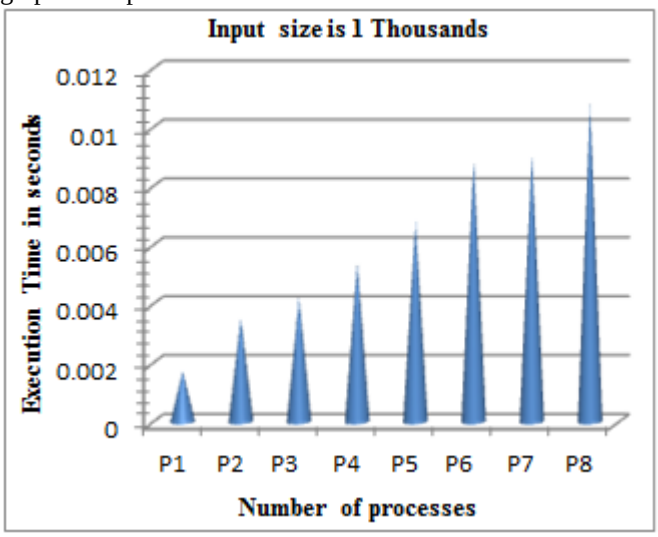


Figure 13.Graphical Representation of HP Cluster (Input size 1 Thousand)

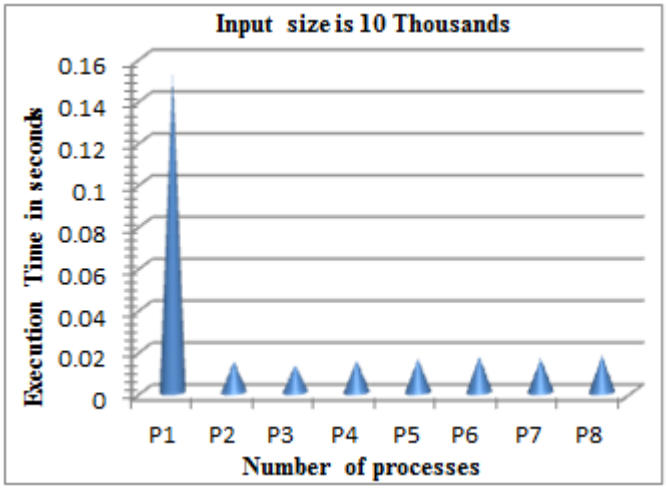


Figure 14.Graphical Representation of HP Cluster (Input size 10 Thousand)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

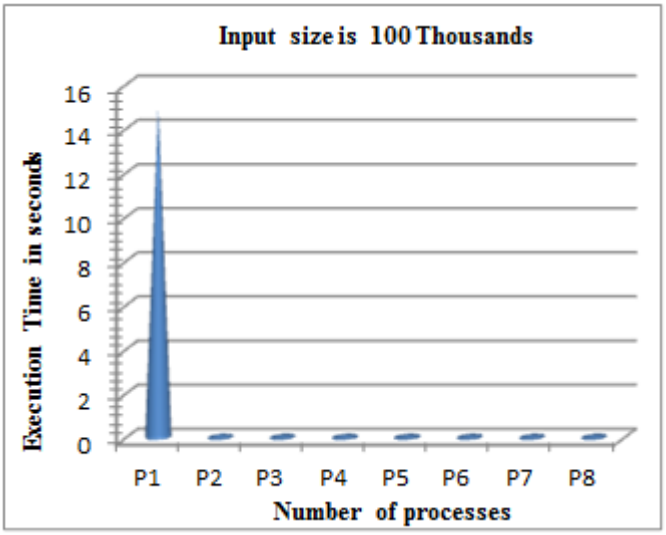


Figure 15.Graphical Representation of HP Cluster (Input size 100 Thousands)

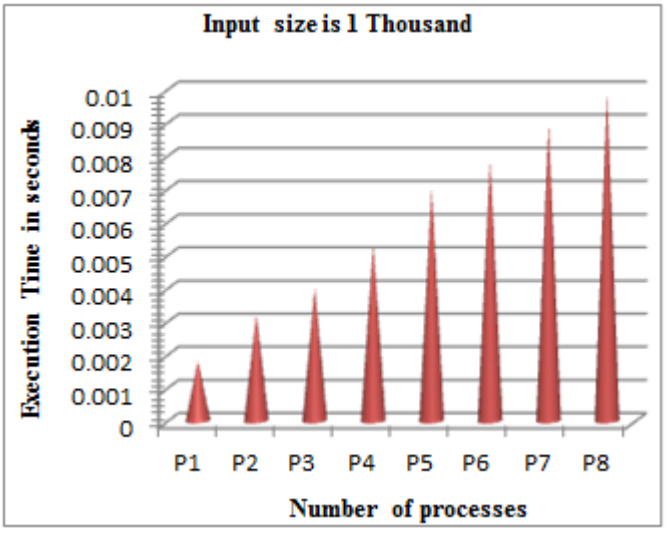


Figure 16.Graphical Representation of DELL Cluster (Input size 1 Thousand)

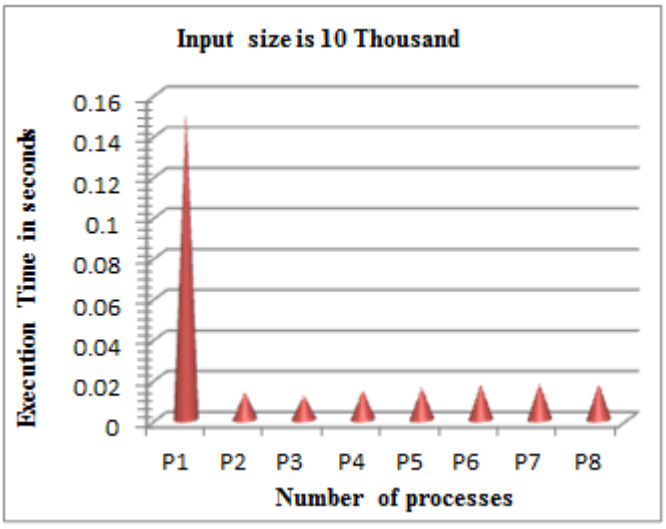


Figure 17.Graphical Representation of DELL Cluster (Input size 10 Thousand)

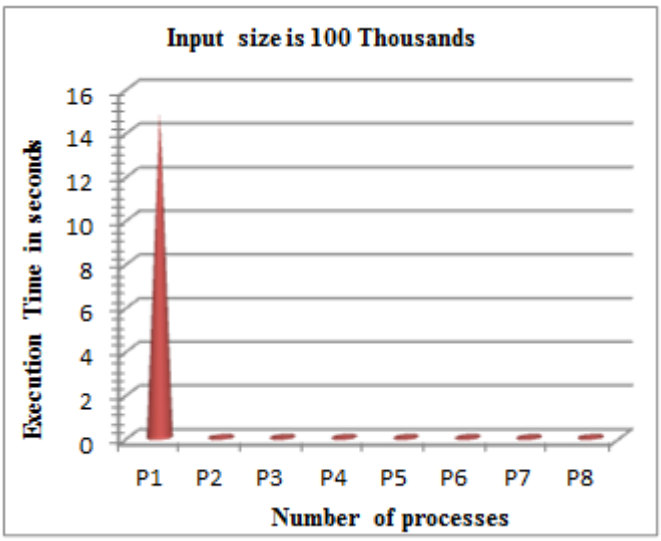


Figure 18.Graphical Representation of DELL Cluster (Input size 100 Thousands)

From above all graphical representations of HP cluster and Dell cluster figures (13 to 18) showing the execution time on different input sizes. Figure 13 shows opposite results from the rest of the charts. It shows that when we execute the same algorithm on multiple processors, it takes longer time than serial execution because the problem size is very small so it doesn't need to send a single problem into multiple processor, single processor can easily handle it. Rest of the charts (from figure 14 to 18) proving that serial execution takes longer time than parallel execution and also as the number of processes increases, the execution time also decreases. As single process (P1) is taking longest execution time whereas P3 is taking shortest execution time amongst all P1 to P8.

1) Time Difference Between Dell and HP Cluster

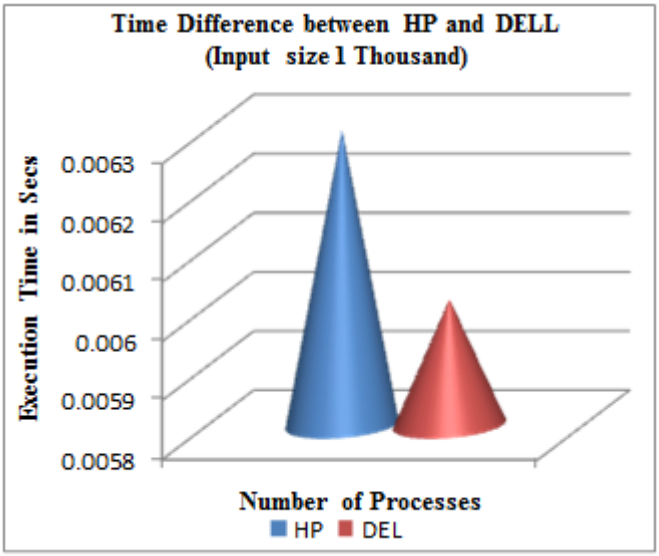


Figure 19.Execution time difference between DELL and HP (Input size 1 Thousand)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

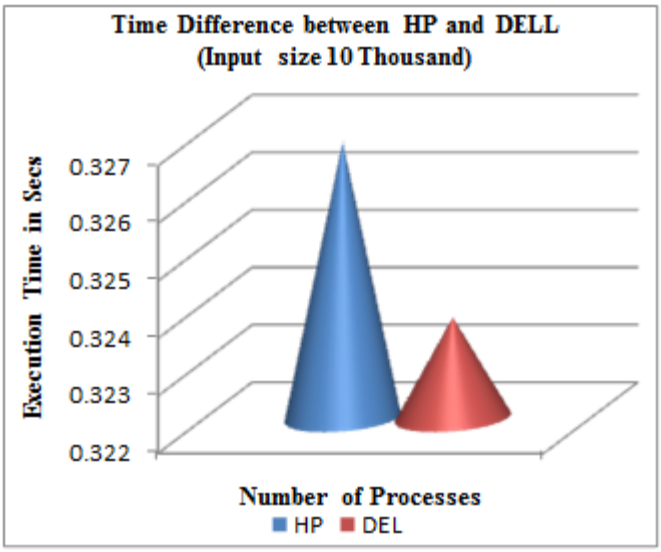


Figure 20.Execution time difference between DELL and HP (Input size 10 Thousand)

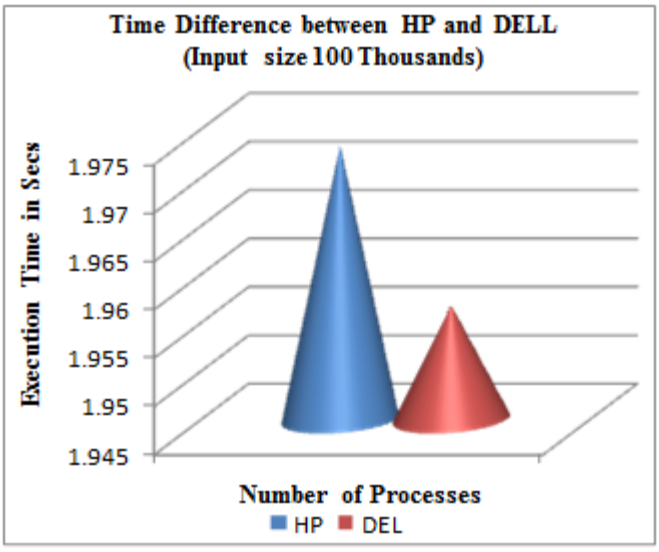


Figure 21.Execution time difference between DELL and HP (Input size 100 Thousands)

It can be observed from above all graphical representation figure (19 to 21), the dell cluster provide better performance than hp cluster. Dell cluster executes the program in minimal time because the size of the problem is very small and it's already discussed in above problem that Dell cluster is suitable for small size problems and all these problems are small in size.

VII. Conclusion

Cluster computing is the most common and popular technology used nowadays to achieve HPC. This research presented two different types of clusters and both have almost same configuration. One cluster is built with two Dell computers and second cluster built with two HP computers. To analyze the performance of these two clusters we executed two different parallel programs on the clusters for pi calculation and quick sort with different problem sizes and compared the results that which cluster executes the algorithm in minimal time. After testing the results and analyzing the performance, we conclude that Dell cluster is suitable for the small size problems because it contains L2 cache only. So it doesn't require shuttling that's why it takes minimum execution time in small size problem. But as the problem size increases, HP cluster provides better performance than Dell cluster because HP contains both L1 cache and L2 cache and Dell has L2 cache only. It means HP cluster suitable for large size problems but unfortunately it is not suitable for the small size problems.

VIII. Future Work

To improve the performance of cluster, the number of nodes can be added up to 4 to 8 or more. The performance of cluster can also be analyzed by using other algorithms. Cluster can also be build using heterogeneous system. Cluster can also be implemented using other operating system like UNIX or MAC etc.

References

[1] J. Tang and C. J. Matyas, "Arc4nix: A cross-platform geospatial analytical library for cluster and cloud computing," *Comput. Geosci.*, vol. 111, pp. 159–166, 2018.

[2] K. Doucet and J. Zhang, "Learning Cluster Computing by Creating a Raspberry Pi Cluster," *Proc. SouthEast Conf. - ACM SE '17*, pp. 191–194, 2017.

[3] N. Kumar and D. P. Vidyarthi, "An Energy Aware Cost Effective Scheduling Framework for Heterogeneous Cluster System," *Futur. Gener. Comput. Syst.*, vol. 71, pp. 73–88, 2017.

[4] J. Eickholt and S. Shrestha, "Teaching Big Data and Cloud Computing with a Physical Cluster," *Proc. 2017 ACM SIGCSE Tech. Symp. Comput. Sci. Educ. - SIGCSE '17*, pp. 177–181, 2017.

[5] A. Mappuji, N. Effendy, M. Mustaghfirin, F. Sondok, R. P. Yuniar, and S. P. Pangesti, "Study of Raspberry Pi 2 quad-core Cortex-A7 CPU cluster as a mini supercomputer," *Proc. 2016 8th Int. Conf. Inf. Technol. Electr. Eng. Empower. Technol. Better Futur. ICITEE 2016*, pp. 7–10, 2017.

[6] J. Kommeri, T. Niemi, and J. K. Nurminen, "Energy efficiency of dynamic management of virtual cluster with heterogeneous hardware," *J. Supercomput.*, vol. 73, no. 5, pp. 1978–2000, 2017.

[7] B. C. Stahl, J. Timmermans, and B. D. Mittelstadt, "The Ethics of Computing," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–38, 2016.

[8] C. K. K. Reddy, K. E. B. Chandrudu, P. R. Anisha, and G. V. S. Raju, "High Performance Computing Cluster System and its Future Aspects in Processing Big Data," *Proc. - 2015 Int. Conf. Comput. Intell. Commun. Networks, CICN 2015*, pp. 881–885, 2016.

[9] A. Rahman, "High Performance Computing Clusters Design & Analysis Using Red Hat Enterprise Linux," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 14, no. 3, pp. 534–542, 2015.

[10] L. Adhianto *et al.*, "HPCTOOLKIT: Tools for performance analysis of optimized parallel programs," *Concurr. Comput. Pract. Exp.*, vol. 22, no. 6, pp. 685–701, 2010.

[11] A. A. Datti, H. A. Umar, and J. Galadanci, "A Beowulf Cluster for Teaching and Learning," *Procedia Comput. Sci.*, vol. 70, pp. 62–68, 2015.

[12] A. Ashari and M. Riasetiawan, "High Performance Computing on Cluster and Multicore Architecture," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 13, no. 4, p. 1408, 2015.

[13] S. Mittal and J. S. Vetter, "A Survey of CPU-GPU Heterogeneous Computing Techniques," *ACM Comput. Surv.*, vol. 47, no. 4, pp. 1–35, 2015.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[14] S. Hasan, A. Ali, and M. A. Y. Al-sa, "Building High Performance Computing Using Beowulf Linux Cluster," vol. 12, no. 4, pp. 1–7, 2014.

[15] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The Cluster Between Internet of Things and Social Networks: Review and Research Challenges," *IEEE Internet Things J.*, vol. 1, no. 3, pp. 206–215, 2014.

[16] K. Banaš and F. Kruzel, "Comparison of Xeon Phi and Kepler GPU performance for finite element numerical integration," *Proc. - 16th IEEE Int. Conf. High Perform. Comput. Commun. HPCC 2014, 11th IEEE Int. Conf. Embed. Softw. Syst. ICESS 2014 6th Int. Symp. Cybersp. Saf. Secur. CSS 2014*, pp. 145–148, 2014.

[17] A. M. Pfalzgraf and J. A. Driscoll, "A low-cost computer cluster for high-performance computing education," *IEEE Int. Conf. Electro Inf. Technol.*, pp. 362–366, 2014.

[18] A. Rumyantsev, "Stabilization of a high performance cluster model," *Int. Congr. Ultra Mod. Telecommun. Control Syst. Work.*, vol. 2015–January, no. January, pp. 518–521, 2015.

[19] I. Journal and M. Physics, "An Introduction To High Performance Computing," no. September 2013, 2014.

[20] G. L. Valentini *et al.*, "An overview of energy efficiency techniques in cluster computing systems," *Cluster Comput.*, vol. 16, no. 1, pp. 3–15, 2013.

[21] P. Petrides, G. Nicolaides, and P. Trancoso, "HPC performance domains on multi-core processors with virtualization," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7179 LNCS, pp. 123–134, 2012.

[22] M. Desai and L. Angeles, "High Performance Computing and Networking," vol. 1823, pp. 1–11, 2000.

[23] D. Ayanda and Y. Adejumo, "A Prototype Model of High Performance Computing Using Beowulf Cluster," vol. 1, no. December, pp. 696–705, 2011.

[24] J. Lee and O. Abuzaghleh, "AC 2011-1301 : IMPLEMENTING AN AFFORDABLE HIGH PERFORMANCE COMPUTING PLATFORM FOR TEACHING-ORIENTED COMPUTER Implementing an Affordable High Performance Computing Platform for Teaching-oriented Computer Science Curriculum," 2011.

[25] N. Sadashiv and S. M. D. Kumar, "Cluster, grid and cloud computing: A detailed comparison," *ICCSE 2011 - 6th Int. Conf. Comput. Sci. Educ. Final Progr. Proc.*, no. Iccse, pp. 477–482, 2011.

[26] D. H. Jones, A. Powell, C. S. Bouganis, and P. Y. K. Cheung, "GPU versus FPGA for high productivity computing," *Proc. - 2010 Int. Conf. F. Program. Log. Appl. FPL 2010*, pp. 119–124, 2010.

[27] W. Haitao and C. Chunqin, "A High Performance Computing Method of Linux Cluster ' s," vol. 2, pp. 83–86, 2009.

[28] V. V. Kindratenko *et al.*, "GPU clusters for high-performance computing," *Proc. - IEEE Int. Conf. Clust. Comput. ICCC*, 2009.

[29] P. Mohammadpour, M. Sharifi, and A. Paikan, "A self-training algorithm for load balancing in cluster computing," *Proc. - 4th Int. Conf. Networked Comput. Adv. Inf. Manag. NCM 2008*, vol. 1, pp. 104–109, 2008.

[30] Y. Tian *et al.*, "High-performance land surface modeling with a Linux cluster," *Comput. Geosci.*, vol. 34, no. 11, pp. 1492–1504, 2008.

[31] D. Göddeke *et al.*, "Exploring weak scalability for FEM calculations on a GPU-enhanced cluster," *Parallel Comput.*, vol. 33, no. 10–11, pp. 685–699, 2007.

[32] L. Chai, Q. Gao, and D. K. Panda, "Understanding the impact of multi-core architecture in cluster computing: A case study with Intel dual-core system," *Proc. - Seventh IEEE Int. Symp. Clust. Comput. Grid, CCGrid 2007*, pp. 471–478, 2007.

[33] A. Nishida, "Building Cost Effective High Performance Computing Environment via PCI Express," *2006 Int. Conf. Parallel Process. Work.*, pp. 519–526, 2006.

[34] C. Engineering and M. M. E. College, "Cluster Computing : A Mobile Code Approach R . B . Patel and Manpreet Singh," vol. 2, no. 10, pp. 798–806, 2006.

[35] C. S. Yeo, R. Buyya, H. Pourreza, R. Eskicioglu, P. Graham, and F. Sommers, "Cluster Computing: High-Performance, High-Availability, and High-Throughput Processing on a Network of Computers," *Handb. Nature-Inspired Innov. Comput. Integr. Class. Model. with Emerg. Technol.*, pp. 521–551, 2006.

[36] Z. Feng, B. Zhou, and J. Shen, "A parallel hierarchical clustering algorithm for PCs cluster system," *Neurocomputing*, vol. 70, no. 4–6, pp. 809–818, 2007.

[37] V. K. Murthy, "High Performance Cluster Computing Using Component-Oriented Distributed Systems," *First Int. Conf. e-Science Grid Comput.*, pp. 522–529, 2005.

[38] P. Charles *et al.*, "X10: An Object-Oriented Approach to Non-Uniform Cluster Computing," *Proc. 20th Annu. ACM SIGPLAN Conf. Object oriented Program. Syst. Lang. Appl. - OOPSLA '05*, vol. 40, no. 10, p. 519, 2005.

[39] Z. Fan, F. Qiu, A. Kaufman, and S. Yoakum-Stover, "GPU cluster for high performance computing," *Proc. ACM/IEEE SC 2004 Conf. Bridg. Communities*, vol. 00, no. 1, 2004.

[40] A. Kuzmin, "Cluster approach to high performance computing," *Comput. Model. New Technol.*, vol. 7, no. 2, pp. 7–15, 2003.

[41] S. Issue and O. H. Computing, "Cluster Computing," vol. 2010, pp. 6–9, 2010.

[42] A. Apon, R. Buyya, H. Jin, and J. Mache, "Cluster Computing in the Classroom : Topics , Guidelines , and Experiences."

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Web Development in Applied Higher Education Course: Towards a Student Self-Regulation Approach

Bareeq A. AlGhannam, Ph.D.
Computer Science and Information Systems Department, College of Business Studies
The Public Authority for Applied Education and Training
Kuwait, Kuwait
ba.alghannam@paaet.edu.kw

Sanaa AlMoumen, Ph.D.
Computer Science and Information Systems Department, College of Business Studies
The Public Authority for Applied Education and Training
Kuwait, Kuwait
sh.almoumen@paaet.edu.kw

Waheeda Almayyan, Ph.D.
Computer Science and Information Systems Department, College of Business Studies
The Public Authority for Applied Education and Training
Kuwait, Kuwait
wi.almayyan@paaet.edu.kw

***Abstract*—Computing is of a complex nature that is rapidly evolving. Applied institutions are specifically guided by their mission to introduce graduates that are ready for the marketplace. Students in computing faculties need to be empowered with self-regulation in able to cope with these changes in technology. A mind shift in both teaching and learning computing need to be acquired to accommodate current and future marketplace demands. This research shows results of a case study that employs the integration of concepts from Cybernetics and Systems Thinking with Project Based Learning (PBL), in the delivery of an introductory web development course. A case study which uses a constructive grounded theory approach, is utilized as an attempt to introduce change, that is part of a curriculum re-evaluation process for a computing major in an applied business college. The results of the research will provide a better understanding of how Systems Thinking and PBL may be introduced to a computing course in hope to promote self-regulation in students.**

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Keywords- Systems Thinking; Computer Supported Learning; Project Based Learning; Constructivism; Applied Higher Education.

I. INTRODUCTION

Teaching and learning computing in applied higher institutions is challenging. Both faculty and students need to comprehend that they are part of an era of rapid changes in technology that is taught in an environment of a special nature; a blend of academic and vocational with a reach-out to the marketplace. Applied course curriculums need to align with their institution's mission, and be able to inject empowered computing graduates that are ready for the marketplace.

Current and future trends in the technology industry need a workforce that are able to cope with rapid change. The computing industry is expanding in an intertwined manner and with a high demand for computing graduates to work in various places. Graduates of applied institutions need to be able to fit into diverse employment requirements, specifically their technical environment. A mind shift in both faculty and students need to be acquired to accommodate the current and future marketplace. To support that process Cybernetics and Systems Thinking needs to be introduced as part of their complex learning. Reference [1] uses a Cybernetics tool called the Viable System Model (VSM) to design a business school curriculum, it also shows an attempt in introducing Systems Thinking into the curriculum to embed sustainability, and presents the benefits of such a practice. Systems Thinking emphasizes the need for holisticnes, and paradigm that involves considerable attention to concepts and principles that emerge into patterns for intervention purposes [2]. It provides a means for better understanding and uncovering unforeseen problems.

Systems Thinking can be supported by the Project Based Learning (PBL) pedagogy. PBL emphasises ownership, self-learning and innovation in participants. Once students acquire such traits, they will have a drive for self improvement and an adjustment to changes in the marketplace. This is called self-regulation in cybernetics. An autonomous state that is viable on any perturbed environment.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

This paper presents a literature review of Applied Higher Education, Systems Thinking and PBL. The literature review emphasizes their concept, and what distinguishes each of them when employed in teaching computing. A case study is used to help understand how such concepts are integrated within a teaching course in web development. The findings will help advance the teaching of applied computing. The case study encapsulates findings from an introductory web development course in the faculty of Computer Science and Information Systems (CSIS) at the Public Authority for Applied Education and Training (PAAET) in Kuwait. The study shows results of the process followed in the course delivery, the instructors' reflection and students feedback. The paper concludes that the integration of concepts from Systems Thinking and PBL prompted the learning experience for both the instructor and students, and helped build initiatives for self regulation in students.

II. LITRERATURE REVIEW

A. *Applied Higher Education*

Applied Higher Education is an educational pedagogy that enables students to either acquire knowledge from the workplace environment, or apply solutions of real work problems in the class. It is distinctive as it lies between theoretical knowledge and vocational training. This makes it challenging for educators to identify the extent at which applied education conveys theory to training, and vice versa.

Literature shows tertiary institutions employing diverse ways of acquiring knowledge from the field. For example, integrated learning, community service learning, project-based learning, field work, virtual simulation, …etc. Each practice empowers students with beneficial skills that could never be acquired through a traditional classroom or lab. Skills that make graduates stand out and compete in the marketplace.

The literature shows that authentic learning, created through simulated experiences, have a positive impact on students' self-efficacy and professional identity development. This approach of learning added a sense of real-world engagement and professional practice [3]. This could be simulated using PBL to embed learners with these characteristics. Reference [4] presents a multi-

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

disciplinary program of a digital information ecology and student-focused praxis, where a curriculum of a collaborative nature learning approach was created and led. That approach has facilitated understanding and connecting experiences, practices and online participation that epitomize a 'new culture of learning'.

Computing curricula in applied pedagogy is complex, because the pace of change in technology is rapid. Teaching a computer science subject that involves more practical hands on training, as well as some theoretical informational guidance, needs to evolve. In order to deliver the knowledge of web-based design, we need to review our teaching method and learning approach. This field of learning is relatively complex, especially when no pre-knowledge practices of how to develop a website design that is usable and well structured. In curriculum design there are course material and teaching design approaches that must be paired to the content of such a complex learning environment. Currently, complex learning became a popular educational approach in many types of inquiry, guided discovery, and particularly in project-based, problem-based and design-based learning materials have not been well defined as empirical research [5].

The basic idea of helping students in complex learning environments, which involves real-life authentic tasks, is to drive teaching, training, and learning to "integrate knowledge, skills, and attitudes, simulate them to learn to coordinate constituent skills, and facilitate transfer of what is learned to new problem situations" [5] [6]. Since real-life website design projects are never a straightforward progression of steps as those defined by Zou and Chan in 2015 [5]. On the other hand, the assessment of educational programs promote the integration of domain skills and must be defined in the same context of the learning process.

It is important to consider "peer" learning. Peer students in groups can sometimes have a bigger role in learning because they are in a similar situation and have the same goals. Unlike teachers or experts, peers do not have power over each other by virtue of their position or responsibilities [7].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Using concepts from Cybernetics would be beneficial as it is widely used to comprehend complex systems. Systems Thinking would aid the process of evaluating and managing such curriculum, a holistic approach that can be used to enhance the delivery of the course.

B. *Cybernetics and Systems Thinking*

Cybernetics is the science that is concerned with communication and control of the man and the machine. It is the science of steersman-ship which is interdisciplinary with diverse branches. Scientists from different fields of mechanical engineering, biology as well as other sciences combined their knowledge to create a greater way for better understanding. Cybernetics was developed by Wiener which was further evolved as a discipline by Ashby, Von Foerster and McCulloch. Scientists worked in a way to complement each other's knowledge. For example, Stafford Beer used Cybernetics in management. He used concepts from human anatomy and how the brain functions, to create models to manage organizations such as in the Viable System Model (VSM). Scholars have made use of the VSM tool and made notes of its difficulties [8], other scholars have called for the sharing of detailed guidance on how such tools should be used to aid novice users [9]. Documentation of case studies employing such tools promotes usage, and encourages others to implement them effectively in diverse investigations.

One of the Cybernetic branches that is developed by Werner Ulrich is Systems Thinking. Systems Thinking is a holistic way in understanding how things unfold over time and space. It is a science that looks at all things as a system [2]. And by that notion we can consider any item, person, or even a conceptual feeling as a system. Once this is done it is essential to define the context boundaries of the system under focus. Boundaries can be vague and the definition relies on the evaluator's intuition and the stakeholders involved. There is a need to identify who is affecting the system under focus, and who the system is having an effect on. The many tools, approaches and methodologies in Systems Thinking provides a means to better understanding in order to uncover unforeseen problems. Reference [10] has gone beyond better understanding by using verbs and actions in Systems Thinking.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The literature shows the importance of Systems Thinking in diverse disciplines and emphasizes the extreme importance of skills needed to succeed as a system thinker. They go further into dividing activities of Systems Thinking into “Gaining Insight” and “Using Insights” [10]. Scholars also viewed Systems Thinking as a system itself, which aided the development of skills from a systems perspective [11].

Educators are aware of the prominence of Systems Thinking [12], and have called for investigating ways to enable learners to a certain level of prophecies [13]. Other scholars have gone further into assessing Systems Thinking as learning outcomes in a university [14]. Learners enrolled in distance learning at the open university are equipped with Systems Thinking to shift their minds and comprehend the socio-technical complexity [12]. This interest of enabling academic institutions with Systems Thinking has also intrigued the computing realm as it provides a means to promote applied learning concepts to students. This will broaden student awareness and ground them to their environment. It also enables students to be aware of changes in the environment, and think of ways to cope with them [15] [16].

c. *Project Based Learning (PBL)*

Academics delve into pedagogies that enhance the learning experiences of students. Throughout the years student learning styles have changed, as well as marketplace demands. Students are required to be equipped with essential qualities that will make competent workers, and able to withstand changes in the environment. Project Based Learning (PBL) is an approach that is capable of providing students with the qualities demanded from the current and future workforce. It is a way to empower students to feel that they are able to make a change [17].

PBL is a student centered approach that engages students to learn [18]. It is widely used by k-12 teachers, specifically primary, as there is evidence to its effectiveness [19]. It is also adopted by various universities and colleges worldwide. PBL is a well structured methodology that is defined clearly. Learners experience a project over a prolonged time period, which makes them cognitively engaged. Literature shows its benefits and drawbacks [20]. It is evident that PBL is known to enhance learning by promoting motivation [20], however, PBL requires specific recommendations such as technology, pedagogy, training, assessments, .. etc [17].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

If computing instructors integrate PBL within their course with the aid of Systems Thinking detailed planning supported by the administration is essential to make change. Literature documents such practices in teaching computing, however they are scarce. There is a need to better understand how PBL could be part of a computing faculty environment, specifically in an applied nature where PBL can provide synergetic results when paired with Systems Thinking.

III. METHODS

A. *Processes and Tools*

A case study gives an in-depth understanding of a specific context [21]. It has been chosen as a tool for investigation in this study, because it fits the goal of the research. The research inquires into the comprehensive nature of how PBL and Systems Thinking is to be integrated within the delivery of an applied computing web development course, and how that would promote self-regulation in students. Context is essential, thus qualitative tools, methods and methodologies are the way to address this research [22] [23], through the utilisation of constructivism paradigm [24].

The project's delivery is empirically documented and derived through the flow of the web development course. The instructor acts as an observer and evaluator. Grounded theory; which is a qualitative paradigm that stresses on the role of the researcher, and the need to be aware of the epistemology and ontology of it [22] was employed.

Reliability and validity in research ensures that the processes followed could be replicated, would give the same results, and that the right tool is used. This is challenging considering the subjective nature of case studies. Diverse scholars have documented detailed processes that are relevant to qualitative research [23]. It requires a validation of a specific nature and rigorous application of the processes ensures that virtue. Rigorousness is enforced throughout the project. It can be seen in defending the ontology and epistemology of the system under focus, triangulation, documentation, ..etc [25]. As for Reliability, this is a more challenging task [25] as it is impossible to get the same results in different subjective contexts as the participants and circumstances could never be replicated.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The instructor in this case study has a background in qualitative research and Systems Thinking, and this qualifies the instructor to be able to follow a rigorous process to encapsulate results that are specific to the context of the environment. The case study is demonstrated below with details of the administration.

B. *Participants and Setting*

The case study is of an introductory web development course. The course is from the faculty of Computer Science and Information Systems (CSIS) department at the College of Business Studies (CBS); one of the colleges of the Public Authority for Applied Education and Training (PAAET) in Kuwait.

PAAET is an educational authority of higher applied education and training. Its operational environment is increasingly complex and resides in a competitive environment. As many other institutions of higher education worldwide, PAAET colleges are facing challenges to respond to national and global market needs, and social change. The CSIS department goal is to increase the proportion of students in this specific discipline. It is vitally important to empower our graduates with workplace attributes by ensuring that the quality of learning in the computing program are both nationally and globally acceptable. In Reference [26], Daniel, assures that despite the substantial uncertainties, the continuing growth of learning analytics must explore the ethical challenges in institutions as a means to drive and shape student support [27], as well as finding opportunities for better and more effective decision making in higher education [28]. Analytics can help learners and instructors recognize danger signs before threats to learning success materialize [29].

Stakeholders involved in the case study are all female -18 students and one instructor- registered in the summer of 2017/2018 academic calender. During the academic year, the course is five credits and six hours/week (four hours theoretical and two hours lab), however during the summer, which is spread over nine weeks, the course is modified to 12 hours/week -three hours/day, four days/week-. All the classes are instructed in the lab. W3schools [30] is an interactive learning website that scholars have categorized as a tutorial-based website [31]. In this study, it is used in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

three ways within the course: First, as an instruction and demonstration method used by the instructor. Second, as a means for application and practice used by the students, and third, as a reference resource during project implementation which is also used by the students.

The main objectives of this course is for the students to develop skills that enable them to design usable front-end Graphic User Interface (GUI) websites. Notepad text editor on the Windows platform is used to write HTML, CSS and basic JavaScript.

Most students find web design course enjoyable and relatively easier than the other courses in the major. This encouraged the authors to add Systems Thinking and PBL as learning tools to this specific course in order to increase student enjoyment, which will better enhance the learning experience.

c. *Procedure and Analysis*

A case study is used to document the delivery of the introductory web development course, through predefined phases that naturally follow the unfolding of the project. Delivery of the course content is not sequential as in the W3schools website and modified accordingly.

The importance of using a specific process for content delivery is for quality purposes. The processes need to be linked with the course content and learning goals. Integration with the context of where the project resides in the community is essential.

Student consent to participate in the study was obtained at the beginning of the course. They were assured that names will not be used in the research, and their reflections on the project will not effect the total grade in any way. The researcher was aware of the ethical issues in qualitative studies and abided by it.

The instructor acts as the facilitator, observer and evaluator. The course delivery through PBL is considered a system, and needs to be analysed accordingly. A background in holistic evaluation enables the instructor to evaluate the system from different granulates, and be able to synthesize unseen problems. An external evaluator is necessary for validating the results.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this case study, the Goal Question Metrics (GQM) questionnaire is used to develop project goals and evaluate them. The questionnaire was administered in a focus group that involved the whole class. The goal was to collect feedback from the students a few days before their final exam.

Also, students were required to individually submit ungraded reflections of problems faced. Patterns of similar problems can be defined more explicitly within the PBL process of delivering the course.

IV. RESULTS

The case study was conducted in a class of 18 female students. Students were encouraged to work in groups of two and three members resulting in a total of five groups. However, four students preferred to work as individuals. The process, tools and procedure described in the methods section was followed. An outline of the grounded process is presented below:

1) *Students were instructed to review local websites and were required as a class to reach consensus on a specific project. Holistic thinking is encouraged by the instructor to widen student perspectives, and involve the students in decision making.*
2) *Students are given the option of working in a group or as individuals. Groups are set with leaders and communication channels are established within the group and between groups. Communication is essential in cybernetics [32].*
3) *Students are required to create a personal log to document all thoughts, work, ideas, and reflections. A grade is given for keeping the log.*
4) *Students are encouraged to find similar webpages to the agreed upon website topic (local and international), and are encouraged to encapsulate common design pattern, synthesize, and analyze the positive and negative features of each (for example layout, font size, colour scheme, infographics, .. etc.)*
5) *Students are required to decide on what information to include in the webpages.*
6) *Students are required to construct the layout design of their webpage on paper, and draft where to place the information.*
7) *Students are required to build their website during class in integral phases according to the topic sequence covered in the class.*
    a. *Instructor inquires the students on their design options, and encourages the discussion of "how to improve the website" throughout the project development.*
    b. *Instructor encourages the interaction between team-members, and between members of other teams during website implementation.*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

c. *Four distinct phases had been observed (Fact Forming, Design Drafting, Initial Building, and Refining). Evaluation is formative after the completion of each phase.*
d. *The flow of the phases is bidirectional and students are encouraged to make modifications in previous phases as needed.*
e. *Students present their work to the class and discussion is encouraged between class members.*

8) *A focus group is formed to collect student feedback on project implementation in the class using GQM.*

V. Discussion

Usually the web design course uses W3schools interactive learning website to deliver its content, practice assignments and aid the students to develop a small project towards the end of the course. In that model of learning, the applied learning experience is achieved at the end of the course. However, in this study applied learning starts to build up from the start of the course because the students get involved with a project from their environment from the beginning of the course, and they evolve as the project evolves throughout the course. By applying this PBL approach, we have achieved higher student engagement throughout the learning experience [18]. In addition, PBL supported student empowerment, and this is noticed when students began to educate their peers in the lab work, and became more confident in doing that as the course progressed. This student engagement and empowerment helped students to meet the projects' deadlines.

Goal Question Metrics (GQM) questionnaire is a method used in software engineering to evaluate metrics that evolved from goals, for the ultimate aim of improving a software or a project. GQM have been used through out the years by diverse practitioners and researchers [33] [34] [35] [36]. The authors employed GQM in this research to initiate discussions facilitated by the instructor within the focus group of students. In this study GQM was used to develop metrics. The metrics that were emphasized were students' enjoyment and empowerment during project implication. There was a consensus that all students enjoyed the project and all participants that made an effort in the project found themselves able to verbally say that "*they are able to complete any web development challenges in the future*".

From the focus group it was noted that students had a better understanding of the subject matter. One student that repeated the course said that "*she finally understood why she is learning these languages*". There was a sense of community between the students within their group and between

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

groups. The collaboration within groups and between different groups promoted interpersonal skills, as well as a sense of responsibility to aid others. Team work in the projects involved many socio-emotional factors that promoted the learning process. Socio-emotional factors also support the development of Computer-Supported Collaborative Learning (CSCL) environment. In [37] the community addressed three main challenges for "adequate incorporation of effective states in CSCL environments: emotional awareness, orchestration of students' interaction and group formation" [37]. In this study W3schools acts as a CSCL and the utilization of PBL supported, by Systems Thinking promotes students to be self-learners, initiative and innovative. These traits have been observed during the administration of the case study. The students developed a pattern of collaboration reflected from their interpersonal behaviour, their personal reflections log and focus groups discussions.

A detailed plan of the project aided the instructor in linking its phases with the learning goals of the course. An explicitly defined rubric for each step in each phase of the PBL project was linked to learning outcomes. The rubric facilitated the process of grading. Moreover, informing students of a bonus for the best project encouraged aspiring students to optimize their project work and aim for excellence from the beginning. Some groups invested in their time where specific groups stayed longer after class and others met before the class.

Results of the case study depends on the unique interaction between students and instructor. Qualitative data provides understandings that could never be encapsulated from quantitative data. Descriptive findings articulate context that represents patterns of behaviour and the processes encapsulated. The first author's reflection is based on individual analysis that is supported by a background in systems thinking.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. Conclusion and Further Studies

The significance of the study lies in introducing and accepting change in computing faculty. A change that includes integration of Systems Thinking and PBL in delivering computing courses.

Instructor evaluation of the course delivery and students feedback are positive which promote the utilisation of PBL in future web development courses. The results show that PBL was delivered through specific process that could be used as a blueprint to follow in future computing courses. This could be extended to other computing courses.

The study also provides better understanding of students' dynamics during PBL in computing courses. The concept of Peer Programming was adopted instinctively between students in the same team. Team collaboration varied depending on the members. This calls for future work that investigates personality types of team members during PBL and its affect on the collaboration, project management and project success.

There is also a need to study incentives and their affect on the quality of the students projects. Extra credits for best work enthusiasts some students personality types to excel. The instructor could employ an external evaluator to choose the best project to ensure that there is no bias.

It is important to note that the results of this research are from a single case study that included all female students. The instructor's reflection is directed by the perspective of self-learning. It is not possible to replicate the dynamics of students and instructor. However, further administration will promote enhancement of the learning process. Further similar case studies of web development is essential to demonstrate the feasibility of applying PBL and Systems Thinking. There is also a need to collect data over time involving both male and female students taking into consideration the gender of the instructor.

For accreditation purposes GQM could be further linked to the program and institutions learning outcomes. Once this is achieved all learning outcomes will be in sync and the course would be autonomous and in self-regulation.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Computing academics try to innovate their teaching, research, and at the same time try to keep up with the rapid changes in technology as well as being involved with the community. Systems Thinking would able these academics to comprehend this complexity. This research calls to instill Systems Thinking through computing programs as skills to be acquired by both instructors and learners.

ACKNOWLEDGMENT

This research is supported by the Public Authority for Applied Education and Training (PAAET) grant research number BS-17-08 in Kuwait.

REFERENCES

[1] A. Gregory and S. Miller, "Using system thinking to educate for sustainability in a business school", *Systems*, vol. 2, 2014, pp. 313-327.
[2] P. Checkland, *Systems Thinking, Systems Practice*. John Wiley, Chichester, UK, 1981, ISBN 0471279110.
[3] M. Merrill, *First Principles of Instruction*. New York: Pfeiffer, 2012.
[4] J. van Merrienboer and P. Kirschner, *Ten Steps to Complex Learning, A systematic Approach to Four-component Instructional Design*, Third Edition, Routledge, Taylor and Francis Group, 2018. Available: http//taylorandfrancis.com. [Accessed September, 16, 2018].
[5] T. Zou and B. Chan, "Developing professional identity through authentic learning experiences", In M. Davis & A. Goody (Eds.), *Research and Development in Higher Education: The Shape of Higher Education*, 2016, vol. 39, pp. 383-391.
[6] J. O'Connell, "Networked participatory online learning design and challenges for academic integrity in higher education", *International Journal for Educational Integrity*, vol. 12, no. 4, 2016, Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).
[7] D. Boud, R. Cohen and J. Sampson, "Peer Learning in higher education, learning from and with each other", *Assessment and Evaluation in Higher Education*, vol. 24, no. 4, pp.413-426, 2013.
[8] M. Orengo, "Theoretical notes regarding the practical application of Stafford Beer's viable system model", *Kybernetes*, vol. 47, no. 2, pp. 262-272, 2018.
[9] S. Hildbrand, and S. Bodhanya, "Guidance on applying the Viable System Model", *Kybernetes*, vol. 44, no. 2, pp. 186-20, 2015.
[10] R. Arnold and J. Wade, "A Complete Set of Systems Thinking Skills ", In Proceedings of *the 27th Annual INCOSE International Symposium (IS 2017)* Adelaide, Australia, July 15-20, 2017. Available https://www.researchgate.net/profile/Ross_Arnold/publication/320246371_A_COMPLETE_SET_OF_SYSTEMS_THINKING_SKILLS/links/59ee04160f7e9b3695758e3d/A-COMPLETE-SET-OF-SYSTEMS-THINKING-SKILLS.pdf, [Accessed September, 17, 2018].
[11] R. Arnold and J. Wade, "A Definition of systems thinking: a systems approach", *Procedia Computer Science*, 2015, doi: 10.1016/j.procs.2015.03.050, Available : https://www.researchgate.net/publication/273894661, [Accessed September, 13, 2018].
[12] M. Wermelinger, H. Jon, L. Rapanotti, L. Barroca, M. Ramage and A. Bandara, "Teaching software systems thinking at the Open University", 2015, In: Proceedings of *The 37th International Conference on Software Engineering*, IEEE, 2015, pp. 307–310.
[13] R. Plate and M. Monroe, "A Structure for Assessing Systems Thinking", In *The 2014 Creative Learning Exchange*, vol.23, no. 2014, pp. 1-12.
[14] M. Monroe, R. Plate and L. Colley, "Assessing an introduction to systems thinking", *Natural Sciences Education*, vol. 44, no. 1, 2015, pp. 11-17. doi:10.4195/nse2014.08.0017.
[15] M. Reynolds and S. Holwell, (eds), *Systems Approaches to Managing Change: A Practical Guide*, Milton Keynes, The Open University in association with Springer-Verlag London Limited, 2010, Available: http://www.open.edu/openlearn/ocw/pluginfile.php/704004/mod_resource/content/6/Introducing-systems-approaches_ch1.pdf, [Accessed September, 13, 2018].
[16] S. Easterbrook, "From Computational Thinking to Systems Thinking: A conceptual toolkit for sustainability computing", In: Proceedings of *The Second information and Communication Technologies ICT4S*, Atlantis Press, 2014, Available http://www.cs.toronto.edu/~sme/papers/2014/Easterbrook-ICT4S-2014.pdf, [Accessed September, 13, 2018].
[17] A. Luis, R. Pedro and M., Ricardo-J, "Project-Based Learning: An Environment to Prepare IT Students for an Industry Career", in *Overcoming Challenges in Software Engineering Education: Delivering Non-Technical Knowledge and Skills*. Yu, Liguo, Eds : IGI Global. p. 230-249, ISBN 978-1466658004, ch012, 2014.

[18] J. Gonzalez, "Project based learning: start Here", *Cult of Pedagogy*, 2016, Available: http://www.cultofpedagogy.com/project-based-learning. [Accessed September, 16, 2018].
[19] B. Barron, D. Schwartz, N. Vye, A. Moore, A. Petrosino, L. Zech, J.D. Bransford, and Technology Group at Vanderbilt, "Doing with understanding: Lessons from research on problem and project-based learning", *Journal of the Learning Sciences*, vol. 7, 1998, pp. 271–311.
[20] P.C. Blumenfeld, E. Soloway, R.W. Marx, J.S. Krajcik, M. Guzdial, and A. Palincsar, "Motivating project-based learning: sustaining the doing, supporting the learning", *Educational Psychologist*, vol. 26, no. 3 and 4, 1991, pp. 369–398.
[21] R. Yin, *Case Study Research: Design and Methods*, 5th ed., Thousand Oaks, CA: Sage, 2014.
[22] S. Khan, "Qualitative research method: grounded theory", *International Journal of Business and Management*, vol. 9, no. 11, 2014, ISSN 1833-3850 E-ISSN 1833-8119, Published by Canadian Center of Science and Education, Available : www.researchgate.com, [Accessed September, 13, 2018].
[23] H. Mohajan. "Qualitative research methodology in social sciences and related subjects", *Journal of Economic Development, Environment and People*, vol. 7, no. 1, 2018, pp. 23-48, Available : www.researchgate.com, [Accessed September, 13, 2018].
[24] H. Lauckner, M. Paterson and T. Krupa, "Using constructivist case study methodology to understand community development processes: proposed methodological questions to guide the research process", *The Qualitative Report*, vol. 17, no. 13, 2012, pp.1-22, Available: https://nsuworks.nova.edu/tqr/vol17/iss13/1, [Accessed September, 13, 2018].
[25] L. Leung. "Validity, reliability, and generalizability in qualitative research.", *Journal of Family Medicine and Primary Care*, vol. 4, no. 3, 2015, pp. 324–327, Available : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4535087/#!po=45.3125, [Accessed September, 16, 2018].
[26] B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges ", *British Journal of Educational Technology*, vol. 46, no. 5, 2015, pp. 904–920.
[27] S. Slade and P. Prinsloo, "Learning analytics: ethical issues and dilemmas", *American Behavioral Scientist*, vol. 57, no. 10, 2013, pp. 1509–1528.
[28] D. Oblinger, "Let's Talk ... Analytics," *EducausE Review*, vol. 47, no. 4, 2012, http://www.educause.edu/library/ ERM1240P.
[29] E. Wagner and P. Ice, "Data changes everything: delivering on the promise of learning analytics in higher education", *EDUCAUSE Review*, vol. 47, no. 4, 2012, pp. 33–42.
[30] W3shools, available: http://www.w3schools.com [Accessed September, 13, 2018].
[31] S. Murthy, A. Figueroa, and S. Rollo, "Toward a large-scale open learning system for data management", in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, ACM, New York, NY, USA, Article 16, pp. 1-4, . doi: https://doi.org/10.1145/3231644.3231673, 2018.
[32] M. Orengo, "Theoretical notes regarding the practical application of Stafford Beer's viable system model", *Kybernetes*, vol. 47, no. 2, pp. 262-272 , 2018.
[33] V. Sylaidis, I. Nanakis, V. Kopanas, "Introducing the Goal-Question-Metric approach to telecommunications software development: the PITA experiment", In: Gritzalis D. (eds) *Reliability, Quality and Safety of Software-Intensive Systems, IFIP — The International Federation for Information Processing*, Springer, Boston, MA, pp. 210-230, 1997.
[34] R.van Solingen and E. Berghout, *The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development*, 2016, McGraw-Hill, England,
Available: https://courses.cs.ut.ee/MTAT.03.243/2015_spring/uploads/Main/GQM_book.pdf, [Accessed September, 13, 2018].
[35] F. Yahya, R. Walters, and G. Wills, "Using Goal-Question-Metric (GQM) approach to assess security in cloud storage", Springer International Publishing AG 2017. V. Chang et al. (Eds.): *Enterprise Security*, LNCS 10131, 2017, pp. 223–240, Available: doi: 10.1007/978-3-319-54380-2_10, Available at http://eprints.soton.ac.uk/411068/1/Using_Goal_Question_Metric_GQM_Approach_to_Assess_Security_in_Cloud_Storage.pdf, [Accessed September, 14, 2018].
[36] B. Lindström, A Software Measurement Case Study using GQM, Master Thesis, Lund Institute of Technology, Amsterdam, 2004. Available http://fileadmin.cs.lth.se/serg/oldsergdok/docsmasterthesis/62_Lindstrom_draft.pdf, [Accessed September, 13, 2018].
[37] R. Reis, S. Isotani, C. Rodriguez, K. Lyra, P. Jaques and I. Bittencourt, "Affective States in Computer-Supported Collaborative learning, Studying the past to drive the future", *Computers and Education*, vol. 120, 2018, pp. 29-50.

AUTHORS PROFILE

Bareeq AlGhannam is an Assistant Professor in the Computer Science and Information Systems Department, College of Business Studies, at the Public Authority for Applied Education and Training in Kuwait. Dr. AlGhannam has a Bachelor of Science in Computer Engineering with a Ph.D. in Software Engineering focused on the realm of Stakeholder Collaboration within software requirements collection. Currently Dr. AlGhannam is conducting various research in Systems Usability Evaluation with emphasis on stakeholders perspectives and on Translations of Standard Usability questionnaires. Additionally, Dr. AlGhannnam is currently investigating research on computing curriculum design/evaluation using cybernetic tools.

Dr Sanaa Almoumen is an assistant professor at the department of Computer Science and Information Systems, College of Business Studies, at the Public Authority for Applied Education and Training in Kuwait. Dr. Almoumen has a PhD in Computer Science from Lancaster University, UK and a Master degree in Computer Science, Systems Software Technology, from University of Sheffield, UK. The research interest of Dr. Almoumen, mainly on the field of Requirements/Software Engineering. Dr. Almoumen is also interested in designing RE approaches and methods for Web based systems, based on cultural and social influences that affect systems design and usability. She has also developed Guidelines on good marketing practices for developing multicultural e-commerce systems. Her current work involves developing innovative teaching and learning processes, for applied computer courses and training.

Waheeda Almayyan is an Associate Professor in the Computer Science and Information Systems Department, College of Business Studies, at the Public Authority for Applied Education and Training in Kuwait. Dr. Almayyan has a PhD in Computer Science from De montfort University, in the UK and a Master degree in Computer Science, Systems Software Technology, from Kuwait University. Dr. Almayyan is mainly interested in the fields of Data Mining , Artificial Intelligence and Usability.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

NNN-C algorithm for Correlated attributes to improve Quality of Data in Distributed Data Mining

S.Urmela*, M.Nandhini
Department of Computer Science, Pondicherry University, India
urmelaindra@gmail.com

Abstract- **To propose a normalization algorithm for DDM, Nearest Neighbour Normalization-Correlation (NNN-C) algorithm to normalize the raw data in local level. At local level, two-level dendrogram is formed by assessing the MAX and MIN values of key attributes (used for classifying records) which is normalized by Nearest Neighbour Normalization. The proposed Nearest Neighbour Normalization-Correlation (NNN-C) algorithm aims to maximize the accuracy and minimize the error rate by efficient, lossless and predicted normalization technique on understanding the data distribution in local levels. In proposed algorithm, the MAX and MIN values of each key-attribute are known prior for each sub-cluster formed in local level leading to efficient normalization and reduced out-of-bound error compared to conventional normalization techniques. Further, the proposed NNN-C algorithm solves problem of inconsistency and redundancy which arises in DDM due to integration of data from several local levels by dendrogram formulation with key-attributes. Experimental implementation on real-world distributed Electronic Health Records (EHRs) and Job Recruitment dataset depicts an improved performance compared to other conventional normalization techniques. Finally, the results are analyzed with conventional normalization techniques: min-max, z-score and decimal scale normalization.**

Keywords: ***Distributed Data Mining, Distributed datasites, Data preprocessing, Data normalization, Nearest Neighbour Normalization-Correlation***

I. INTRODUCTION

Raw data contains unwanted noise, redundant, missing and inconsistent data. Data preprocessing techniques are applied to make the data more noise-free, irredundant for mining/pattern extraction. If there are much noisy and redundant data present while mining, then resultant knowledge discovery task becomes a tedious process. Data quality affects the DM task. Data preprocessing is one of the crucial step in any DM process. Data preprocessing methods include, data cleaning, data integration, data transformation and data reduction[1].

Data cleaning is the process of cleaning noisy data, filling missing data and solving data duplication. Data integration combines data-items from multiple independent sources into a single dataset. Data reduction is helpful in working with minimized or reduced dataset to obtain knowledge with good quality without compromising integrity of raw data. In data transformation, the raw data are transformed into form suitable for mining process[2]. Data preprocessing methods followed for DM cant' be applied for Distributed Data Mining (DDM) because heterogeneous raw data in several local levels with different entities need to be preprocessed. Further, the processed data from several local levels need to be brought together at global level for mining[2].

Normalization is the process of preprocessing the data to fall within a certain range which reduces data inconsistency. Normalization technique developed for DM is not suitable for DDM since the technique is devised for centralized environment and data distribution is not considered. A suitable normalization algorithm needs to be applied on raw data at each local level before mining[3]. Further, for privacy-preserving of data in local and global levels, normalization of data is a method to enhance the data privacy and result accuracy[4].

[2]proposed decimal scale normalization for DDM which transforms the raw data from original interval to specified interval. It fits the normalized data in specified range and out-of-bound error is triggered when the normalized data fall out of the specified range. The proposed work lacks in understanding the data distribution (max and min values of attribute) for each local level and outlier values are not normalized.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[5]compared k-means clustering algorithm performance with min-max, z-score and decimal scaling normalization for DDM. Out of the three normalization methods, decimal scaling normalization out-performs former two normalization methods. The proposed decimal scaling normalization for k-means clustering algorithm lacks in data distribution understanding utilizing common max and min value for all the clusters.

In this paper, a normalization algorithm for Distributed Data Mining (DDM), Nearest Neighbour Normalization-Correlation (NNN-C) algorithm is proposed which normalizes data by forming clusters of similar characteristics (based on key-attributes of records). A two-level dendrogram or more (based on key-attributes of records) is formulated at each local level. For each level, different key-attributes are used. At each level, a branch of dendrogram represents a cluster. For the clusters retrieved at each level, key attributes MAX value is normalized. This way of classifying records in two-level with inter-dependency is called correlation. The MAX and MIN values of each key-attribute are known prior for each sub-cluster formed in local level leading to efficient normalization, reduced out-of-bound error on understanding the data distribution in local level and two-level or more (depending on key-attribute) dendrogram (correlation technique) leads to accurate result retrieval. The problem of inconsistency and redundancy arising in DDM is solved by dendrogram formulation with key-attributes. Dendrogram is formulated with different key-attributes corresponding to each local level.

The organization of paper is as follows: Section 2 discusses data preprocessing methods and data normalization methodologies. Section 3 depicts proposed NNN-C algorithm for effective retrieval of records in distributed environment. Section 4 presents the datasets considered for evaluation and performance analysis of proposed algorithm. Section 5 summarizes the paper.

II. DATA NORMALIZATION METHODOLOGIES

Data normalization is an approach of data transformation which scales attributes to fall within a specified range[5]. It reduces data inconsistency and leads to efficient memory and space utilization. Data transformation is a part of data preprocessing tasks. Data preprocessing includes, data cleaning, data integration, data transformation and data reduction[6].

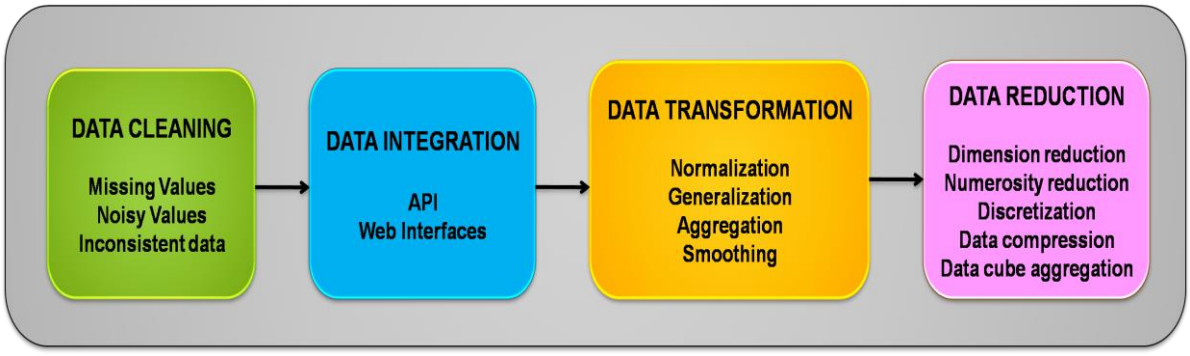


Figure 1. Data Preprocessing methods.

Data cleaning is the process of filtering noisy data (outlier values or erroneous data), missing data (incomplete dataset or aggregate data), inconsistent (containing discrepancies) and data duplication. Dirty data can lead to improper mining process. Filling missing values or ignoring inconsistent values are solutions for data cleaning. Data integration combines data-items from multiple independent sources into a single dataset. Data transformation transforms raw data into form suitable for mining process. Data reduction is helpful in working with minimized or reduced dataset to obtain knowledge with good quality without compromising integrity of raw data[14]. Table I depicts data preprocessing classification and its description.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table I

Data Preprocessing Classification and Description.

Data preprocessing	Classification	Description
Data cleaning	**Missing values[6]**	Ignoring the data Replacing the data Filling data manually
	Noisy values[6]	Clustering process – outlier values are smoothened Binning process - data is smoothened by sorting the values Regression process – linear regression for replacement
	Inconsistent values[7]	Manual correction or with help of knowledge engineering tools/preprocessing routines.
Data integration	**Originating source integration[7]**	Combines data from individual distributed data sources.
	Communication integration[8]	Integration by web interfaces or through Application Program Interface (API)
	Data type integration[9]	Integration by storing the same attribute value with different data type
	Data value integration[10]	Integration by using different representation for same attribute value
Data transformation	**Normalization[11]**	Attribute value is scaled to fall within a specified range
	Generalization[11]	Replaced with higher data by higher concept hierarchies
	Aggregation[12]	Smaller data are aggregated to form large data
	Smoothing[12]	Removes noise values from the data
Data reduction	**Dimension reduction[13]**	Irrelevant or weak attributes are removed
	Numerosity reduction[13]	Data replaced by smaller data representations
	Discretization[13]	Replaced by higher values using concept hierarchies
	Data compression[13]	Encoding methods are used to reduce data size
	Data cube aggregation[13]	Aggregation method applied to raw data in constructing granules of data cube

A lot of approaches have been proposed in recent years to improve the accuracy of DDM by using optimized normalization methods. Description of normalization methodology is discussed in table II[15].

Min-Max Normalization: linear transformation on raw data is done in min-max normalization. It is a simple methodology which probably fit the raw data in a defined range. It transforms the raw data from an existing interval to defined interval[16].

$$Z' = Z - \frac{\text{min value of b} * (c-d) + d}{\text{Max value of b} - \text{min value of b}}$$

where , Z – raw data, Z' – normalized data

Defined range [c,d]

z-score Normalization: transforms by converting the raw data values to a common scale with zero average value and standard deviation as one[17].

$$Z' = Z - \frac{\text{mean}}{\text{standard deviation}}$$

where Z – raw data, Z' – normalized data

Decimal Scaling Normalization: transforms by converting the raw data values on moving the decimal values of considered feature[18].

$$Z' = \frac{Z}{10^j}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

where Z – raw data, Z' – normalized data, j – smaller integer value Max (|D'|) < 1

Table II

Data Normalization Methods

Data normalization	Merits	Demerits
Min-max normalization[16]	Simple normalization technique Probably fits the normalized data in specified range	No outlier value normalization Not suitable for DDM/universal DDM No out-of-bound trigger Lacks knowledge of data distribution
z-score normalization[17]	Fits the normalized data in specified range	No outlier value normalization Not suitable for universal DDM Lacks knowledge of data distribution
Decimal scale normalization[18]	Fits the normalized data in specified range Out-of-bound error is triggered when the normalized data fall out of the specified range	No outlier value normalization Not suitable for universal DDM MAX and MIN value of key-attributes needs to be known prior Lacks knowledge of data distribution

In order to overcome the shortcomings of above discussed normalization methodologies in DDM, Nearest Neighbour Normalization-Correlation (NNN-C) algorithm is proposed which normalizes data by understanding the data distribution in each local level without prior knowledge of MAX and MIN values of key-attributes (used for classifying records). The proposed algorithm works well with universal datasets leading to effective and efficient data retrieval. By forming two or more level of dendrogram (correlation technique), even outlier values are normalized leading to accurate result retrieval.

III. NEAREST NEIGHBOUR NORMALIZATION CORRELATION (NNN-C) ALGORITHM

The proposed Nearest Neighbour Normalization-Correlation (NNN-C) algorithm includes two levels of data processing,

- Filling missing values of raw data
- Dendrogram formulation by Nearest Neighbour Normalization-Correlation (NNN-C)

A. Filling missing values of raw data

In proposed work, the data structure used for storage of heterogeneous and homogeneous distributed data sites (horizontally and vertically partitioned data) is multi-linked list. Consider EHRs dataset, the data storage at local distributed data site is depicted in fig. 2. At each distributed data sites, the missing values are filled from the relative data of corresponding entity. For any type application, there are two types of filling missing values. Filling demographic data and filling application-specific data. Demographic data includes filling general information. Example: filling entity age from DOB, etc. Application-specific data includes filling application-oriented data[19]. For medical dataset, clinical results of patient are filled by relative computation of corresponding data. For example, say in hypertension dataset if the value of SBP (Systolic Blood Pressure)/DBP (Diastolic Blood Pressure) for a continuous range is 121/76, 134/74, 152/78 then the next value is computed as relative mean (136/76)[20].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

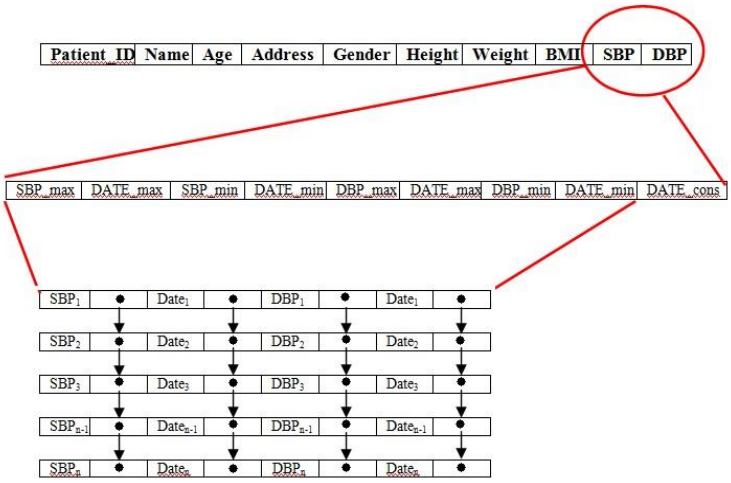


Figure 2. Data Structure representation

Say, for a patient at each local level, if SBP/DBP value is monitored daily, it leads to huge volume of data values. To overcome the problem of maintaining history of records, for every 20 data values of SBP and DBP database is consolidated. Max and min value of SBP and DBP among the old 20 records along with time recorded of max and min values and the time data was consolidated is noted.

B. Dendrogram formulation by Nearest Neighbour Normalization-Correlation (NNN-C)

Consider EHRs dataset, in fig. 2 SBP and DBP values are stored as multi-linked list data. From the clinical data, MAX and MIN values of key-attributes are calculated using which clusters are formulated for mining. For example if SBP value for a single entity reading is 120,134,115,132,134,121,117,128 then MAX = 134 and MIN = 115. Similarly DBP value reading is 80,79,83,82,82,84,82,80 then MAX = 84 and MIN = 79. At each local level, attributes name is different. ICD-9 code (International statistical code for Classification of Diseases)[21] is used as a standard reference for classifying patients diagnosed with disease.

Consider the case of classifying patient diagnosed with hypertension disease. There are two-levels involved in formulating dendrogram. At first level, the dendrogram is formulated with SBP values as shown in fig.3. According to ICD-9 code, the range of classifying SBP is <=120, 121-140,141-160,161-180,181-200 and >200 on analyzing the (MAX,MIN) pair. From the fig.3 on first level classification clusters are formed say C1 to C24 with defined range. In C1, patients (P1-P10) fall under category with SBP <=120. From the (MAX,MIN) pair of 10 patients, all 10 MAX value is normalized to largest MAX value in corresponding cluster. Likewise for all 24 clusters, same normalization technique is followed.

At second level, from the normalized clusters second level of classifying patients with DBP is done. According to ICD-9 code, the range of classifying DBP is <=80, 81-100,101-120, >120 on analyzing the (MAX,MIN) pair. The way of classifying records in two-way as discussed above is called correlation. Correlation among two variables means they are interdependent on each other. For diagnosing a patient with hypertension, both SBP and DBP values are monitored. Hence the algorithm proposed is called NNN-C algorithm. From the fig.3 on second level classification sub-clusters with defined range of DBP is formed. In C11, patients (P2, P5 and P6) fall under category with DBP <=80. From the (MAX,MIN) pair of those three patients, all three MAX value is normalized to largest MAX value of DBP in corresponding cluster. Likewise for all sub-clusters, same normalization technique is followed. According to the example in fig. 3, 96 sub-clusters are formed by NNN-C which in-depth classifies patients with hypertension disease.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

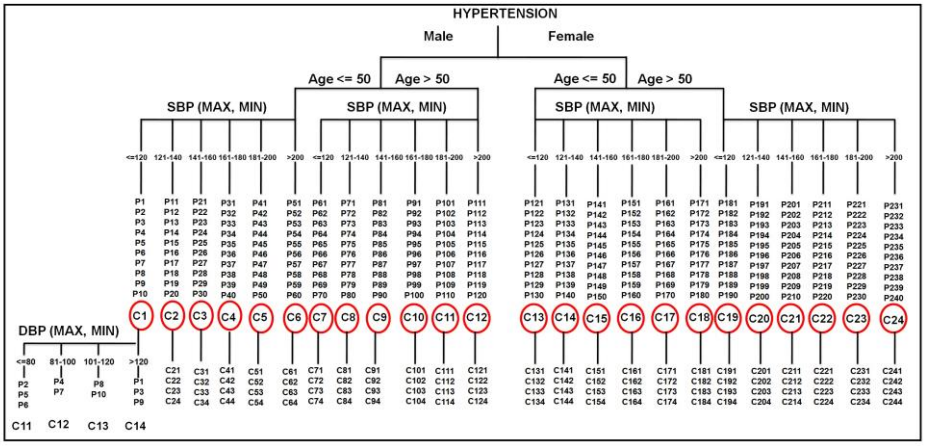


Figure 3. Dendrogram formulation - Nearest Neighbour Normalization-Correlation

Fig. 4 depicts correlation and normalization on EHRs for a single cluster C1 retrieved from dendrogram. Initially cluster set is retrieved on analyzing the (MAX,MIN) pair of SBP value. After forming cluster with SBP value, MAX value is normalized for all the clusters formed. Then DBP value is analyzed which forms sub-cluster with (MAX,MIN) pair of DBP value. Consider for C1, four sub-clusters are formed. Sub-clusters formed for C1 are C11,C12,C13,C14. After sub-clusters formation with SBP and DBP, MAX value of DBP in each cluster is normalized with highest MAX value of corresponding cluster set.

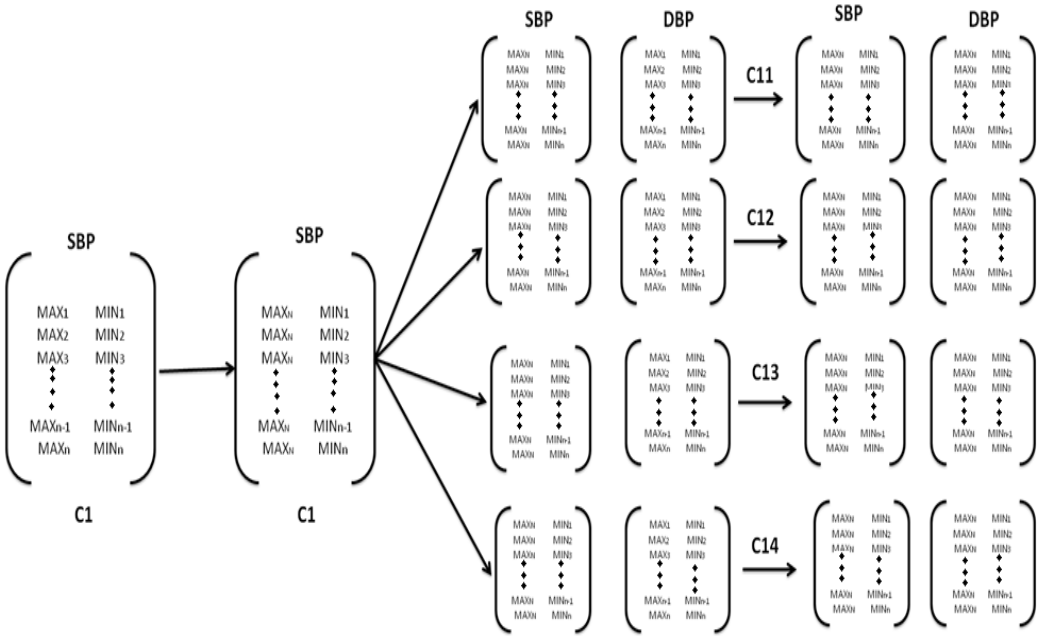


Figure 4. Correlation and Normalization on EHRs – Hypertension

The performance of proposed NNN-C algorithm is evaluated with two applications: EHRs and job recruitment dataset. Performance of proposed algorithm discussed in next section is evaluated with effectiveness and normalization evaluation metrics and compared with min-max, z-score and decimal scale normalization techniques.

IV. EMPIRICAL ANALYSIS

The experiment is carried out on a single 64-bit machine, windows 7 OS having 3GHz Intel dual core processor with 4GB main memory. The proposed algorithm is coded in C# and implemented in HDFS (Hadoop Distributed File

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

System) distributed environment. The proposed NNN-C algorithm is implemented with seven distributed local levels both for EHRs and Job Recruitment dataset. Seven local levels are created with a server at global level. Comparative analyses have been carried out with performance evaluation metrics of proposed NNN-C algorithm with conventional normalization approaches. The proposed normalization-correlation algorithm is implemented on openly-available real-world EHRs and Job Recruitment obtained from UCI repository[21].

A. Case study 1: Electronic Health Records

EHRs are becoming worldwide popular way of maintaining health records of a person. The difference between EHR and Patient Health Record(PHR) is that PHR includes medical record for only a particular disease diagnosed for a person (for a short time period say, 3 years) whereas EHR includes medical record of a person right from their birth till death (for lifetime). EHR includes person therapeutic history, drug usage, allergies and test outcomes. It is supported by EHR management and decision support. In DDM, EHRs are utilized and updated in either homogeneously/heterogeneously distributed local level. Scenario of EHRs model is shown in fig 5.

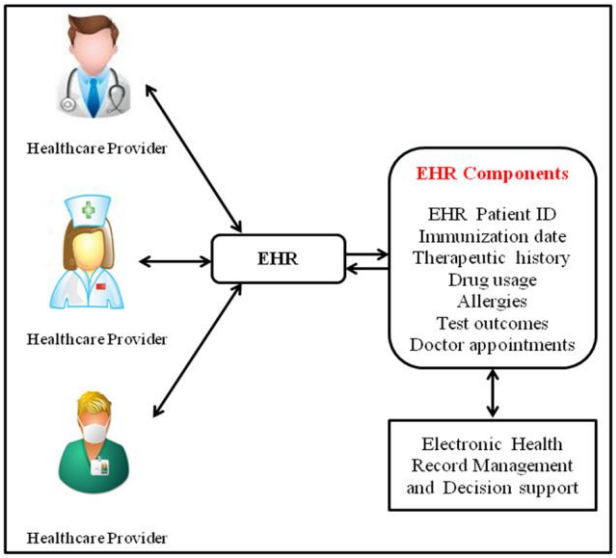


Figure 5. Electronic Health Record Working Model

B. Case study 2: Job Recruitment Dataset

For each candidate, their profiles collectively include the following information:

- **Personal information** such as candidate name, age, gender, and their communication address
- **Education information** consists of candidate undergraduate, postgraduate and higher-graduate specialized subjects and its corresponding grades**.** Scenario of job recruitment network is shown in fig 6.

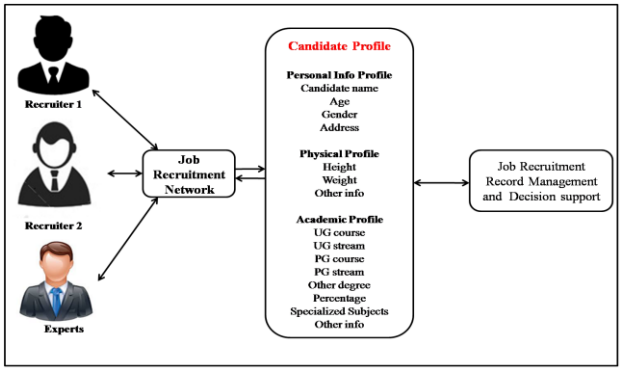


Figure 6. Job Recruitment network model

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C. Analysis of case studies

Both the case studies considered are analyzed on different parameters as shown in table III: stakeholders, expert systems used, local and global level DDM users, no. of local levels considered for evaluation of proposed algorithm, profiles of case study with commonality, common attributes and key attributes of applications used for formulating dendrogram.

Table III

Comparison of Case Study 1: EHRs and Case Study 2: Job Recruitment

Parameters	Electronic Health Records (EHRs)	Job Recruitment records
Stakeholders	Health care providers (doctors, nurses, etc)	Recruiters
	Patients	Candidates
	Health experts	Job recruitment experts
Expert systems	EHR management and decision support	Job Recruitment record management and decision support
DDM-Local level users	Health care personnel	University academic personnel
DDM-Global level users	Health care personnel/other government personnel related to Health	Job recruiter personnel
Profiles	Personal info profile	Personal info profile
	Anthropometric profile	Physical profile
	Clinical results profile	Academic profile
	Medication/allergies vaccination profile	Extra certification profile
Common attributes	Patient Name Age Address Gender	Candidate Name Age Address Gender
Key-attributes	Hypertension: SBP DBP Age Diabetes: GlyHb (Glycalated Hemoglobin) Age	Job recruitment: Qualification Experience Add-on courses Age
Functionality	Adding new EHR record at local level	Adding new candidate record at local level
	Updating existing EHR record at local level	Updating existing candidate record at local level
	Deleting EHR record at local level	Deleting candidate record local level
	Filtering EHR record based on user query at global level	Filtering candidate record based on user query at global level

D. Performance Evaluation Metrics

Performance of proposed NNN-C algorithm is analyzed on effectiveness metrics: accuracy, precision, recall and F-measure and normalization technique evaluation metrics: Mean Squared Error (MSE) and Root Mean Square Error (RMSE)[4].

E. Effectiveness Metrics

Accuracy represents a measure of effective retrieval of records based on user query under normalization algorithm. The mathematical formulation is expressed as[22],

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\textbf{Accuracy (\%)} = \frac{TP + TN}{\text{Total Records}}$$

where, TP – True Positive, TN – True Negative

Precision represents a measure that the retrieved normalized result is relevant to user query. The mathematical formulation is expressed as[22],

$$\textbf{Precision (P)} = \frac{TP}{TP + FP}$$

where, TP – True Positive, FP – False Positive

Recall represents a measure that only relevant normalized records corresponding to user query are retrieved. The mathematical formulation is expressed as[22],

$$\textbf{Recall (R)} = \frac{TP}{TP + FN}$$

where, TP – True Positive, FN – False Negative

F-measure represents normalized harmonic mean measure of precision and recall. The mathematical formulation is expressed as[22],

$$\textbf{F-measure (F)} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

F. Normalization technique evaluation Metrics

Mean Squared Error is defined as the difference between predicted data and actual data. The mathematical formulation is expressed as[5],

$$\textbf{MSE} = \frac{1}{n} \sum (i = 1 \text{ to } n) (X_i - X_i')^2$$

where, n – number of predictions, X_i – original data, X_i' – normalized data

Root Mean Square Error is defined as the square root of difference between predicted data and actual data. The mathematical formulation is expressed as[5],

$$\textbf{RMSE} = \sqrt{\text{MSE}}$$

G. Performance Analysis

The proposed NNN-C algorithm is evaluated with state-of-art normalization approaches: min-max normalization[16], z-score normalization[17] and decimal scaling normalization[18] along with proposed NNN without Correlation. NNN without correlation formulates only one-level of dendrogram irrespective of number of key-attributes in considered application (EHRs and Job Recruitment). Performance analysis of proposed NNN-C algorithm is analyzed with accuracy, precision, recall, F-measure, MSE and RMSE metrics.

Table IV

Effectiveness Measures Evaluation

Dataset	Metrics (%)	Min-max normalized data	z-score normalized data	Decimal scaling normalized data	NNN without C normalized data	NNN-C normalized data
EHRs	**P**	0.6942	0.6418	0.9012	0.9045	0.9434
	R	0.6571	0.6002	0.8449	0.8756	0.9023
	F	0.6751	0.6203	0.8721	0.8937	0.9224

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Job Recruitment	**P**	0.6582	0.6590	0.8390	0.8759	0.9381
	R	0.6337	0.6381	0.8054	0.8571	0.9010
	F	0.6482	0.6476	0.8189	0.8668	0.9218

Table IV depicts effectiveness measures comparison of proposed NNN-C algorithm with the state-of-art normalization techniques. Proposed algorithm exhibits more precision, recall and F-measure (94.34%, 90.23%, 92.24% for EHRs and 93.81%, 90.10%, 92.18% for job recruitment) compared to other normalization techniques because the distribution of original data in local levels is considered in proposed NNN-C algorithm. Min-max normalization (69.42%, 65.71%, 67.51% for EHRs and 65.82%, 63.37%, 64.82% for job recruitment) shows better performance compared to z-score (64.18%, 60.02%, 62.03% for EHRs and 65.90%, 63.81%, 64.76% for job recruitment) because min-max technique fits the data in specified range and an out-of-bound error is triggered when normalized value deviate from the specified range. In z-score, no out-of-bound error or outlier value normalization is present. Decimal scale normalization (90.12%, 84.49%, 87.21% for EHRs and 83.90%, 80.54%, 81.89% for job recruitment) exhibits equal performance as proposed algorithm but an attributes maximum and minimum values need to be known prior. NNN without correlation normalization (90.45%, 87.56%, 89.37% for EHRs and 87.59%, 85.71%, 86.68% for job recruitment) exhibits less equal values compared to proposed algorithm but more than conventional approaches because values of MAX and MIN is known prior. In conventional techniques, data distribution at each local level and global level is not considered while converting the raw data to normalized data.

In proposed NNN-C algorithm, dendrogram formulation (EHRs (Hypertension): key-attributes- SBP and DBP, job recruitment (software tester): key-attributes- qualification, and experience) of sub-levels based on MAX and MIN values of key attributes leads to efficient result prediction based on user query. The more the levels of dendrogram, efficient is the result retrieval.

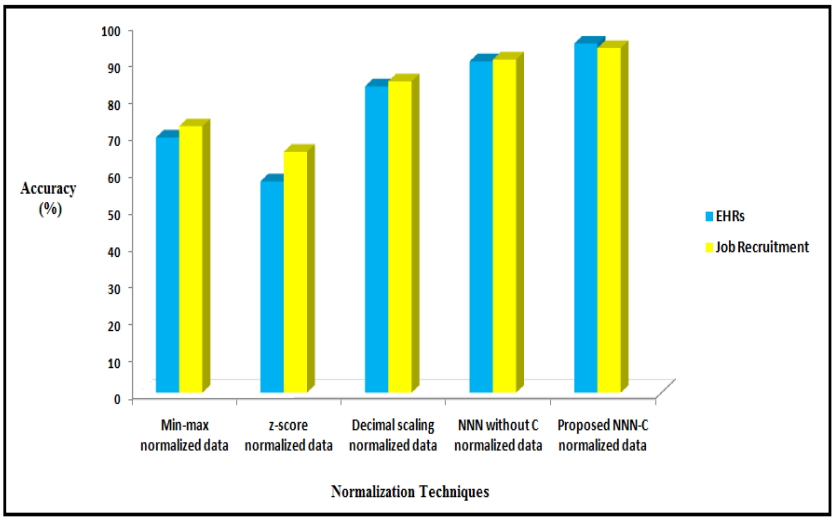


Figure 7. Accuracy Vs Normalization techniques

Fig. 7 depicts accuracy comparison of proposed NNN-C algorithm with the state-of-art normalization techniques. Proposed algorithm exhibits more accuracy (94.81% for EHRs and 92.46% for job recruitment) compared other normalization techniques: min-max normalization (69.28% for EHRs and 72.58% for job recruitment), z-score (57.39% for EHRs and 65.34% job recruitment) and decimal scaling (83.20% for EHRs and 85.28% for job recruitment). In decimal scaling, outlier values or out-of-bound data is considered (pre-knowledge) and normalized within specified range whereas in z-score and min-max normalization knowledge of original attribute range is not known to define the range. NNN without correlation normalization (89.52% for EHRs and 90.17% for job recruitment) exhibits

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

more accuracy than former three conventional normalization techniques because result prediction is efficient by dendrogram formation of sub-levels. The more sub-levels of dendrogram based on MAX and MIN of key-attributes leads to efficient result accuracy. In proposed NNN-C algorithm, outlier values are normalized and pre-knowledge of attributes maximum and minimum value is not necessary.

Fig. 8 depicts misclassification errors comparison of proposed NNN-C algorithm with the state-of-art normalization techniques. Min-max exhibits more error rate (57.82% for EHRs and 65.73% for job recruitment) because boundary value is not known compared to z-score (53.69% for EHRs and 64.71% for job recruitment) and decimal scaling (42.17% for EHRs and 44.79% for job recruitment). Proposed NNN-C algorithm (29.31% for EHRs and 34.81% for job recruitment) exhibits better performance than min-max and z-score because of reduced out-of-bound error. Since, decimal scaled data considers outlier values misclassification errors is less than former two techniques. NNN without C algorithm (35.74% for EHRs and 41.27% for job recruitment) exhibits less error rate compared to NNN-C but more compared to former three conventional techniques because of not considering correlation which classifies dataset in two-levels avoiding misclassification errors. Proposed NNN-C algorithm considers all the data within specified range and outlier values for normalization.

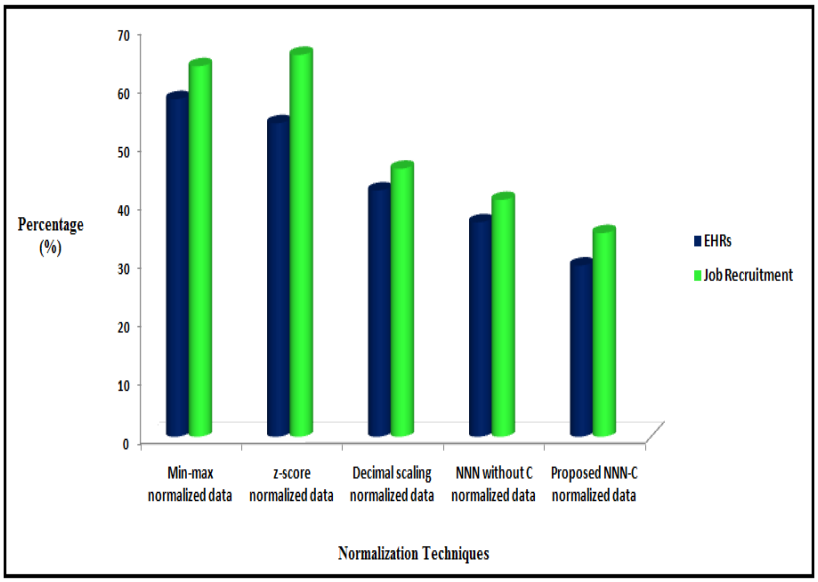


Figure 8. Misclassification Errors Vs Normalization techniques

Table V depicts MSE and RMSE comparison of proposed NNN-C algorithm with the state-of-art normalization techniques. Min-max exhibits less MSE and RMSE (1.2313, 1.1782 for EHRs and 1.3482, 1.1611 for job recruitment) compared to z-score (1.3621, 1.1892 for EHRs and 1.4573, 1.2071 for job recruitment) because out-of-bound error is triggered in min-max normalization when the normalized range doesn't fall within the specified range. In decimal scaling MSE and RMSE (1.1182, 1.2781 for EHRs and 1.1683, 1.0808 for job recruitment) error rate is less compared to former two normalization techniques because prior knowledge of max and min value of normalized attribute is known. NNN without C algorithm (0.9045, 0.9510 for EHRs and 1.7493, 1.3226 for job recruitment) exhibits more error rate compared to proposed NNN-C because data distribution is not considered leading to less accuracy rate. Proposed NNN-C algorithm (0.1972, 0.3958 for EHRs and 0.2034, 0.4509 for job recruitment) exhibits better performance than former three state-of-art normalization techniques because data distribution at local levels is considered minimizing the error rate. Proposed NNN-C algorithm considers all the data within specified range and outlier values for normalization.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table V

(MSE, RMSE) Vs Normalization Techniques

Dataset	Metrics (%)	Min-max normalized data	z-score normalized data	Decimal scaling normalized data	NNN without C normalized data	NNN-C normalized data
EHRs	**MSE**	1.2313	1.3621	1.1182	0.9045	0.1972
	RMSE	1.1782	1.1892	1.2781	0.9510	0.3958
Job Recruitment	**MSE**	1.3482	1.4573	1.1683	1.7493	0.2034
	RMSE	1.1611	1.2071	1.0808	1.3226	0.4509

V. CONCLUSION

In this paper, a normalization-correlation method based on nearest neighbour cluster formation is proposed. Experimental implementation on EHRs and Job Recruitment datasets shows that the NNN-C algorithm is effective for normalization technique in DDM. Performance is evaluated with effectiveness and normalization technique evaluation metrics and compared with state-of-art normalization techniques: min-max, z-score and decimal scale normalization. In proposed algorithm, the MAX and MIN values of each key-attribute are known prior for each sub-cluster formed in local level leading to efficient normalization and reduced out-of-bound error compared to conventional normalization techniques.

REFERENCES

[1] Alfredo Cuzzocrea, "Models and algorithms for high-performance distributed DM", Elsevier Journal of Parallel and Distributed computing, 2013, Vol.73, No.93, pp.281-283.
[2] Alfredo Cuzzocrea, "Models and algorithms for high-performance data mining", Elsevier Journal of Parallel and Distributed computing, 2013, Vol.73, No.92, pp.271-273.
[3] L. Al-Shalabi, "Coding and Normalization: The Effect of Accuracy, Simplicity, and training time", RCED'05, Al-Hussain Bin Talal University, 2006, Accepted.
[4] L .Polkowski, S .Tsumoto, T. Lin, "Rough Set Methods and Applications", Physica-Verlag, Heidelberg New York, 2016, pp.49-88.
[5] K. Cios, W. Pedrycz, R. Swiniarski, "Data Mining Methods for Knowledge Discovery", Kluwer Academic Publishers, 1998.
[6] S. V. S. Ganga Devi, "A Survey on Distributed DM and its Trends", International Journal of Research in Engineering & Technology (IJRET), 2014, Vol.2, No.3, pp.107-120.
[7] S.Gopal Krishna Patro, Pragyan Parimita Sahoo, Ipsita Panda, Kishore Kumar Sahu, "Technical Analysis on Financial Forecasting", International Journal of Computer Sciences and Engineering, 2017, Vol.3, No.6, pp.1-6.
[8] G .Gora, A .Wojna, "RIONA: A New Classification System Combining Rule Induction and Instance-Base Learning", Fundamenta Informaticae, 2012, Vol.51, No.4, pp.369-390.
[9] Gora, A. Wojna, "RIONA: A Classifier Combining Rule Induction and k-NN Method with Automated Selection of Optimal Neighbourhood", Proceedings of the Thirteenth European Conference on Machine Learning, ECML, Helsinki, Finland , Lecture Notes in Artificial Intelligence, 430 Springer-Verlag, 2016, pp.111-123.
[10] Han, M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, USA, 2001.
[11] G.Manikandan, N.Sairam, S.Sharmili, S.Venkatakrishnan, "Data Masking – A few new Techniques", International conference on research and development prospects on engineering and technology (ICRDPET -2013), E.G.S Pillay Engineering college, Nagapattinam, march 29-30, 2013.
[12] C.J. Mertz, P.M. Murph, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, 1996.
[13] A.O. Ogunde, O. Folorunso, A.S. Sodiya, "A partition enhanced mining algorithm for distributed association rule mining systems", Egyptian Informatics Journal, 2015, Vol.16, No.3, pp.297-307.
[14] R. Quinlan, "Induction of Decision Trees", In Machine Learning. Kluwer Academic Publishers, 1986, Vol.1, pp.81-106.
[15] Sanjaya K. Panda, Subhrajit Nag, Prasanta K. Jana, "A Smoothing Based Task Scheduling Algorithm for Heterogeneous Multi-Cloud Environment", 3rd IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, Waknaghat, 11th - 13th Dec 2017.
[16] Sanjaya K. Panda, Prasanta K. Jana, "Efficient Task Scheduling Algorithms for Heterogeneous Multi-cloud Environment", The Journal of Supercomputing, Springer, 2017.
[17] Shalabi, L.A., Z. Shaaban, B. Kasasbeh, "Data Mining: A Preprocessing Engine", J. Comput. Sci., 2017, Vol.2, pp.735-739.
[18] Tsoumakas, G., Spyromitros-Xioufis, J Vilcek, E., Vlahavas, I, "Distributed Data Mining", In Proceedings ECML/PKDD, Workshop on Mining Multidimensional Data (MMD'08), 2008, pp.30-44.
[19] Vinaya Sawant, Ketan Shah, "A review of Distributed DM using agents", International Journal of Advanced Technology & Engineering Research (IJATER), 2013, Vol.3, No.5, pp.27-33.
[20] M .Wojnarski, "LTF-C: Architecture, Training Algorithm and Applications of New Neural Classifier", Fundamenta Informaticae, 2009, Vol.54, No.1, pp.89-105.
[21] YanLi, ChangxinBai, ChandanK.Redd, "A distributed ensemble approach for mining health care data under privacy constraints", Journal of Information Sciences, Vol.330, 2018, pp.245-259.
[22] S.Urmela, M.Nandhini, "CBPLSA - An effective Collaborative Filtering Algorithm for Distributed Data Mining on Electronic Health Records", The Official Journal of Institute of Integrative Omies and Applied Biotechnology (IIOAB Journal), 2018, Vol 9, No.3, pp. 01-08.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Clustering of patients for prediction of glucose levels based on their glucose history

Claudia Margarita Lara Rendon
División de Estudios de Posgrado e Investigación
Instituto Tecnológico de León
León, Guanajuato, México
claudia.rendon@itleon.edu.mx

David Asael Gutiérrez Hernández
División de Estudios de Posgrado e Investigación
Instituto Tecnológico de León
León, Guanajuato, México
david.gutierrez@itleon.edu.mx

Marco Antonio Escobar Acebedo
Departamento de Universidad de la Salle
Universidad de la Salle
León, Guanajuato, México
marcoaescobar@gmail.com

Raúl Santiago Montero
División de Estudios de Posgrado e Investigación
Instituto Tecnológico de León
León, Guanajuato, México

Carlos Lino Ramírez
División de Estudios de Posgrado e Investigación
Instituto Tecnológico de León
León, Guanajuato, México

Manuel Ornelas Rodríguez
División de Estudios de Posgrado e Investigación
Instituto Tecnológico de León
León, Guanajuato, México

***Abstract*— The present work makes the prediction of glucose levels in 70 subjects doing use of the Empirical mode decomposition (EMD) and model ARIMA, with which the prediction of the glucose levels was made, later was used K-means for clustering of the subjects according to its errors of prediction, with base in that we can say that if subject the account with an historical minor to the 30 data, is not possible the prediction of glucose levels with model ARIMA because the prediction error is of the 87,54%. However, to obtain the prediction of 2 glucose levels with an error RMSE under it is required of a minimum of 110 data, with which we obtain an error of the 61,59%, nevertheless, between the rank of 40 to 100 data the error average is in a 66%. For the case of the predictions of 3, 4 and 5 data, it was observed that the prediction counts on 53,3% of error if a minimum of 150 data is had, nevertheless, was observed that between more IMF they are obtained from the historical one better is the prediction.**

Keywords- ARIMA; EMD; IMF; K-means; Prediction

I. INTRODUCTION

The diabetes mellitus, is a chronic disease that takes place when elevated glucose levels occur in blood because the organism lets produces or does not produces enough the denominated hormone insulin, or does not manage to use this hormone of effective way. And based on the information that provided International Diabetes Federation (IDF) in his eighth edition, the criteria of diabetes diagnosis have struggled and they have been updated throughout decades, but, according to the present criteria of Organizacion Mundial de la Salud (OMS), diabetes by means of the observation of elevated glucose levels in blood is diagnosed [1].

The glucose levels are a factor important for the control of the diabetes, nevertheless, for foretelling the glucose levels is not a simple task, which we can verify that when making a continuous monitoring of the glucose levels of each person obtains a series of time nonlinear, if to this problem is added to him that the glucose can be altered by variables that cannot be controlled, this makes that the task of predicting this type of information is complicated still more, for that reason the objective of this work is to be able to give a classification with base in the historical one of glucose of the subjects with diabetes, for the prediction of the levels of glucose.

It is known that the patients with diabetes must control continuously their glucose levels in blood and fit the doses of insulin, looking for to maintain the glucose levels in the ranks normal. Kevin Plis in the 2014 proposes an automatic model of prediction that she warns to the people of changes in his glucose levels so that she allows them to take preventive measures. Based on the article of Plis the model surpasses to the experts in diabetes to predict the glucose levels and could be used to anticipate almost one fourth part of the hypoglycemic events with 30 minutes of anticipation, nevertheless the corresponding precision is single of 42%, and most of the false alarms they are in regions near hypoglycemia reason why it says that the patients who can respond to these alert not will see harmed by intervention [2].

Nevertheless, in the last years a method has been developed to analyze data nonlinear and nonstationary, with this method is made a empirical model decomposition (EMD), which allows that any complicated data set can be disturbed in a finite number of intrinsic mode functions (IMF). This method of decomposition is adaptive and, therefore, highly efficient. Since the decomposition is based on the local time scale characteristic of the data, he is applicable to processes nonlinear and nonstationary [3]. The EMD is a method of signal processing, proposed by Huang in 1998 [4], is used to deal with data nonstationary and nonlinear, is intuitive, direct, and a posteriori adaptive, with the base of the decomposition based and derived from data [3].

The aim of this model EMD is to soft the series of time, and to discompose the original signal in several IMF, which is a

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

linear technique, to represent of adaptive way the nonstationary signals like sum of components[4].

Zhongda, Chengcheng, Gang and Yi, proposed a method of prediction based on the EMD, with the purpose of improving the precision of the prediction of the series of time, since through model EMD the series of time can be discompose in different components from frequency, giving passage to the components after the decomposition eliminate the long, prominent correlation and the different local characteristics from the series of time, that can reduce the nonstationary series of time. The results that they obtained after the simulation showed that the prediction method is efficient [4].

Noemi Nava, Tiziana Di Matteo and Tomaso Aste [5], worked with a methodology of prediction on great scale for temporary series nonlinear and nonstationary on the base of a combination of EMD and Support Vector Regression (SVR). Where the proposal was to discompose the series with EMD to obtain a finite set of IMF and a residual, giving like result that the IMF are more adapted to the prediction of the series of time original. In this work each IMF and the residual were foretold separately, and continued with the reconstruct the series of time with the sum of the predicted components. With this work it was concluded that the use of the EMD model can improve the forecasts, however the limited improvement for the short term horizons may be due to the border effects of the EMD, which produce oscillations at the ends of the MFIs and disturb the first anticipation steps.

Xueheng Qiu along with their collaborators presented a method made up of algorithm EMD and the approach of deep learning [6], where I am made the decomposition of the series in several IMF and making use of a network of deep learning. The obtained results of the prediction of each IMF were added to obtain a value added for the demand of load. The obtained results of the simulation showed that the proposed method is quite favorable is compared with other methods of forecast.

In the area of the health as in other fields the amount of data available has become more difficult to manipulate and to analyze to obtain information. It is for that reason that the known algorithm as K-means is one of the more popular methods of grouping for massive data sets. In the work of Marco Capóa, Aritz Perez, Jose A. The Lozanos in 2017, propose an efficient approach to the K-means problem destined to massive data [7].

II. METHODOLOGY

A. EMD & IMF

The method EMD is Known as empirical model decomposition, is used to work with data nonstationary and nonlinear, this method is intuitive, direct, adaptive, and a posteriori, with the base of the decomposition based and derived from the data. The decomposition is based on the simple supposition that any data is formed by different simple intrinsic ways of oscillation. Each mode, can or not to be linear, will have he himself number of ends and crossings by zero. Each one of these oscillating ways is represented by Intrinsic Mode Function (IMF).

A IMF is defined as a function that fulfills the following requirements: a) In all the data set, the number of ends and the number of crossings by zero must at the most be equal or different by one. b) In any point, the average value of the surrounding one defined by the maximum premises and the surrounding one defined by the local minimums.[8]

B. ARIMA

Processes ARIMA are a class of stochastics processes that are used to analyze series of time, propose by Box and Jenkins [9].

With models ARIMA it is possible to be predicted the future values of a series of time from the historical behavior, with no need to identify the underlying factors in the movements of the variable in the time. For this reason, it has known them like nonstructural models. To identify the model appropriate for each series of time it is required to know the integration degree of the series to predict, reason why it is necessary to have, the amount of times that must be differentiated a series until obtaining a stationary progression. In this mode, it is said that a series follows a process ARIMA (p, r, q) where p represent the autoregressive terms, q of average moving respectively and r represents the degree of integration [10].

C. K-means

Clustering is used for problems of grouping, where the entrance is a set of points with a notion of similarity between each pair of points, and a parameter k, that specifies the wished number of groups. The purpose is to group the points in k clusters so that the points assigned to him himself cluster are similar. A form to obtain this partition is to select to a set of k centers and assign each point to the most popular objectives is the cost function k-means, that diminishes the sum of square distances between each point and its center. In the Euclidian space, this is equivalent to diminish the variance of the points assigned to him himself group [11].

The K-means algorithm has been identified like one of the 10 main algorithms in mining of data [7].

III. RESULTS

We worked with a data base that counts on a total of 70 subjects of test [12], of this base the glucose levels fasting were selected of each one of the subjects. Each analyzed series account with a N number of samples, indicating the glucose

Identify applicable sponsor/s here. *(sponsors)*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

levels, each used series has a different behavior, this caused because each subject has its habits.

Each one of the series passed through the decomposition of model EMD to obtain the finite set of IMF of each one of the series, with base in the number of IMF of each one of the series was make the predictions with model ARIMA making use of the IMF, terminating the prediction, we continue whit the classification use the algorithm of k-means with the purpose was applied of classifying the signals in 2, 3, 4, 5 and 6 groups.

With model ARIMA was made the prediction of 2, 3, 4 and 5 data forward, the results seemed not to be favorable, can be observed in the table 1 as the error obtained average when calculating the average quadratic error of each one of the obtained predictions of the 70 signals.

TABLE I. PREDICTION ERROR BY TOTAL NUMBER OF PREDICTIONS

Prediction	2	3	4	5
Error	77.6867	83.2172	83.0212	87.8855

With the predictions obtained in the case of 2 data forward, we know that the error average is of 77,68% and with base in the made tests of classification with K-means where K had the values from 2 to 6, the best classification obtained is the one that are in figure 1, it which shows 4 groups, in where group 2 it counts on the error average of 61,63% and subjects classified in this group count on the characteristics that the historical registry this between the 80 and 110 data, and count on 5 and 6 IMF.

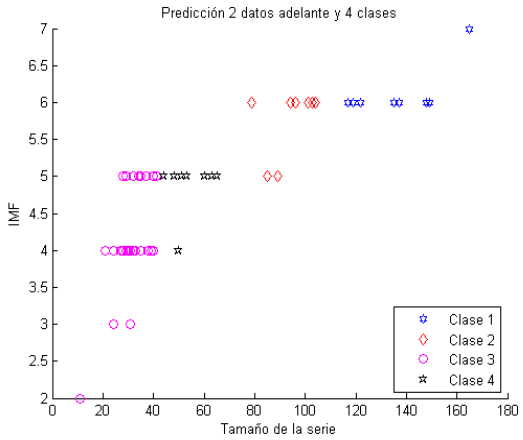


Figure 1. Clustering with K-means of the prediction of 2 data in the 70 test subjects, K = 4

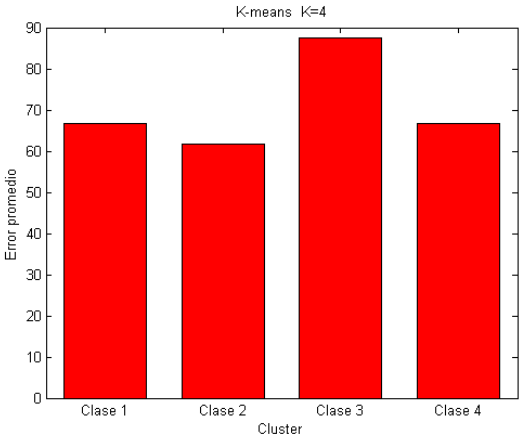


Figure 2. Average error per cluster obtained from the prediction of 2 glucose levels per test subject

The error average of each class can be observed in the table 2, which shows: the class, the total of subjects classified in each class and the error average of each class.

TABLE II. AVERAGE ERROR PER CLUSTER, IN THE PREDICTION OF 2 DATA

Cluster	Total subjects	Average error (RMSE)
1	11	66.5985
2	10	61.6390
3	40	87.5495
4	9	66.5985

In the case of the prediction of 3 data forward these were the obtained results, in figure 3 can be observed the clusters obtained with the method of K-means where K= 3, with this value could be observed the error was less for the prediction of 3 data.

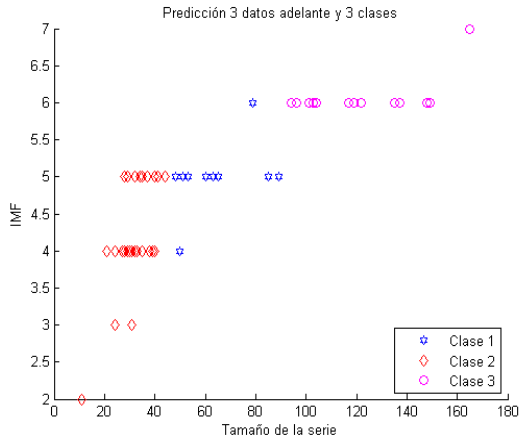


Figure 3. Clustering with K-means of the prediction of 3 data in the 70 test subjects, K = 3

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

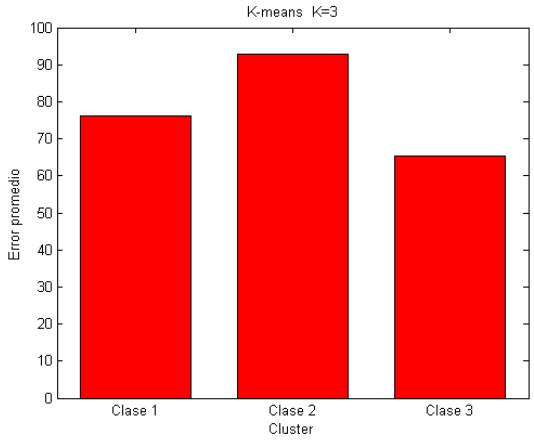


Figure 4. Average error per cluster obtained from the prediction of 2 glucose levels per test subject

In table 3 we can observe the error average by cluster, the total of subjects that conform cluster and the number of cluster obtained.

TABLE III. AVERAGE ERROR PER CLUSTER, IN THE PREDICTION OF 3 DATA

Cluster	Total subjects	Average error (RMSE)
1	21	61.7139
2	49	92.1528

In the case of the prediction of 4 data forward these were the obtained results, in figure 5 can be observed the clusters obtained with the method of K-means where K= 2, with this value could be observed the error was less for the prediction of 4 data.

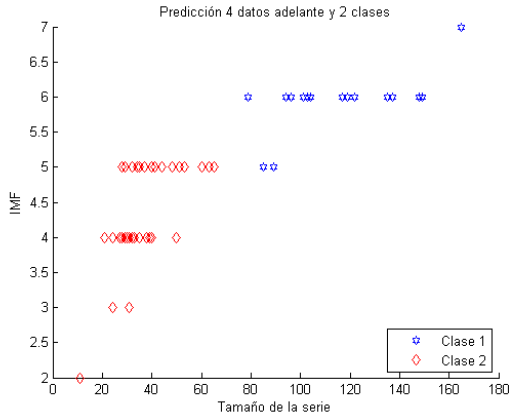


Figure 5. Clustering with K-means of the prediction of 4 data in the 70 test subjects, K = 2

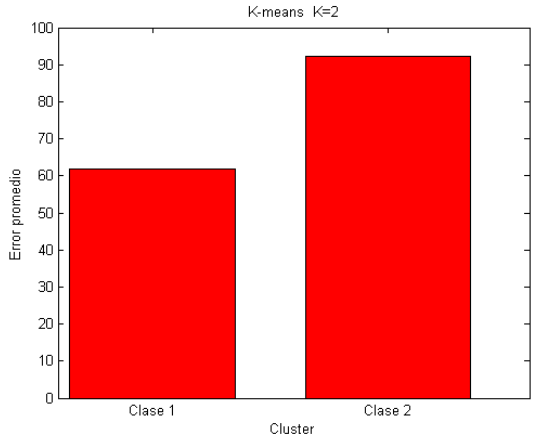


Figure 6. Average error per cluster obtained from the prediction of 4 glucose levels per test subject

The error average of each class can be observed in the table 4, which shows: the class, the total of subjects classified in each class and the error average of each class

TABLE IV. AVERAGE ERROR PER CLUSTER, IN THE PREDICTION OF 4 DATA

Cluster	Total subjects	Average error (RMSE)
1	11	76.2937
2	41	92.8531
3	18	65.4994

In the case of the prediction of 5 data forward these were the obtained results, in figure 7 can be observed the clusters obtained with the method of K-means where K= 5, with this value could be observed the error was less for the prediction of 5 data.

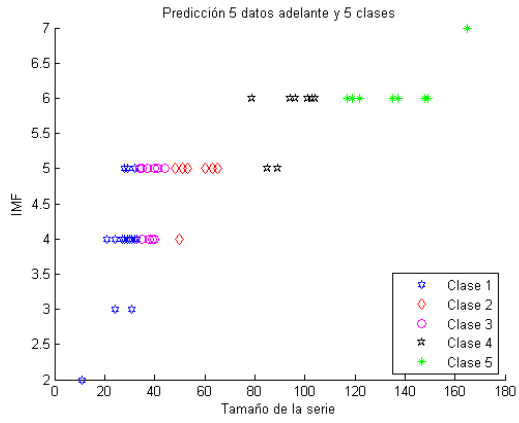


Figure 7. Clustering with K-means of the prediction of 5 data in the 70 test subjects, K = 5

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

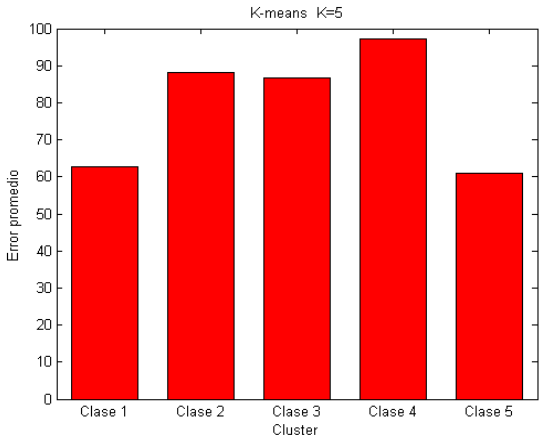


Figure 8. Average error per cluster obtained from the prediction of 5 glucose levels per test subject

The error average of each class can be observed in the table 5, which shows: the class, the total of subjects classified in each class and the error average of each class

TABLE V. AVERAGE ERROR PER CLUSTER, IN THE PREDICTION OF 5 DATA

Cluster	Total subjects	Average error (RMSE)
1	28	103.0453
2	8	95.0658
3	13	83.5242
4	10	71.7670
5	11	63.8821

Until this point we can to observed that in the classifications of 4, 5 and 6 he himself group of subjects selects itself, that agree with the characteristics that have an historical glucose registry in a rank of 80 to 110 data and that when happening the series through EMD obtains 5 or 6 IMF, will have in average an error of the 61,6390%, therefore can be said that to classify in 4 groups it is sufficient to define to the possible candidates for the prediction of its glucose levels.

IV. CONCLUSION

In order to conclude the analysis it is possible to be said that the candidates for the prediction of glucose levels will be those that count on historical registry with 80 or more glucose samples, and that when happening these data through the decomposition of model EMD obtain a minimum of 5 IMF, this to can prediction of 2 data forward, for the case of looking for to have a prediction of 3, 4 or 5 data forward will be with an historical registry with 90 or more glucose samples, and that when pass these data through the decomposition of model EMD obtain a minimum of 6 IMF.

ACKNOWLEDGMENT *(HEADING 5)*

The authors wish to thank CONACYT, the National Technological Institute of Mexico and the Technological Institute of León for their support of this work..

REFERENCES

[1] International Diabetes Federation, *IDF Diabetes Atlas Eighth Edition 2017*. 2017.

[2] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management," *Mod. Artifical Intell. Heal. Anal.*, pp. 35–39, 2014.

[3] N. E. Huang, "New method for nonlinear and nonstationary time series analysis: empirical mode decomposition and Hilbert spectral analysis," *Proc. SPIE*, vol. 4056, no. 301, pp. 197–209, 2000.

[4] T. Zhongda, M. Chengcheng, W. Gang, and R. Yi, "Approach for time series prediction based on empirical mode decomposition and extreme learning machine," *2018 Chinese Control Decis. Conf.*, pp. 3119–3123, 2018.

[5] N. Nava, T. Matteo, and T. Aste, "Financial Time Series Forecasting Using Empirical Mode Decomposition and Support Vector Regression," *Risks*, vol. 6, no. 1, p. 7, 2018.

[6] X. Qiu, Y. Ren, P. N. Suganthan, and G. A. J. Amaratunga, "Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting," *Appl. Soft Comput. J.*, vol. 54, pp. 246–255, 2017.

[7] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," *Knowledge-Based Syst.*, vol. 117, pp. 56–69, 2017.

[8] G. Rilling, P. Flandrin, and P. Gonçalvès, "ON EMPIRICAL MODE DECOMPOSITION AND ITS ALGORITHMS."

[9] J. Contreras, R. Espínola, F. J. Nogales, and A. J. Conejo, "ARIMA models to predict next-day electricity prices," *IEEE Trans. Power Syst.*, vol. 18, no. 3, pp. 1014–1020, 2003.

[10] D. R. Broz and V. N. Viego, "Precios De Productos Almacenables," *Madera y Bosques*, vol. 20, no. Primavera 2014, pp. 37–46, 2014.

[11] S. Gupta, R. Kumar, K. Lu, B. Moseley, and S. Vassilvitskii, "Local search methods for k-means with outliers," *Proc. VLDB Endow.*, vol. 10, no. 7, pp. 757–768, 2017.

[12] M. Michael Kahn, MD, PhD, Washington University, St. Louis, "UCI Machine Learning Repository: Diabetes Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes. [Accessed: 01-Oct-2017].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2018-2019
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/